

University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

UNIVERSITY OF SOUTHAMPTON

Saliency for Image Description and Retrieval

by

Jonathon Stephen Hare

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

April 2006

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Jonathon Stephen Hare

We live in a world where we are surrounded by ever increasing numbers of images. More often than not, these images have very little metadata by which they can be indexed and searched. In order to avoid *information overload*, techniques need to be developed to enable these image collections to be searched by their content.

Much of the previous work on image retrieval has used global features such as colour and texture to describe the content of the image. However, these global features are insufficient to accurately describe the image content when different parts of the image have different characteristics. This thesis initially discusses how this problem can be circumvented by using salient interest regions to select the areas of the image that are most interesting and generate local descriptors to describe the image characteristics in that region. The thesis discusses a number of different saliency detectors that are suitable for robust retrieval purposes and performs a comparison between a number of these region detectors. The thesis then discusses how salient regions can be used for image retrieval using a number of techniques, but most importantly, two techniques inspired from the field of textual information retrieval.

Using these robust retrieval techniques, a new paradigm in image retrieval is discussed, whereby the retrieval takes place on a mobile device using a query image captured by a built-in camera. This paradigm is demonstrated in the context of an art gallery, in which the device can be used to find more information about particular images.

The final chapter of the thesis discusses some approaches to bridging the semantic gap in image retrieval. The chapter explores ways in which un-annotated image collections can be searched by keyword. Two techniques are discussed; the first explicitly attempts to automatically annotate the un-annotated images so that the automatically applied annotations can be used for searching. The second approach does not try to explicitly annotate images, but rather, through the use of linear algebra, it attempts to create a *semantic space* in which images and keywords are positioned such that images are *close* to the keywords that represent them within the space.

Contents

Nomenclature	ix
Acknowledgements	xi
1 Introduction	1
1.1 Aims and Objectives	2
1.2 Contributions	2
1.3 Thesis Structure	3
2 Background	5
2.1 Content-based Retrieval	5
2.1.1 Textual Information Retrieval	6
2.1.1.1 Classical Text Retrieval: The Vector Space Model	6
2.1.1.2 The Vector Space Model Extended: Latent Semantic Indexing	9
2.1.2 Image Retrieval	11
2.1.2.1 Applications of Image Retrieval	12
2.1.2.2 Retrieval Paradigms	13
Browsing.	13
Navigation.	13
Query By Example.	14
Query By Sketch.	14
2.1.2.3 The Fundamental Bases of CBIR	14
Feature Extraction.	14
High Dimensional Indexing.	14
Image Retrieval System Design.	14
2.1.2.4 Recognition and retrieval using salient interest points	15
2.1.3 Retrieval Evaluation	16
2.1.3.1 Precision and Recall	16
2.1.3.2 Single Value Summaries	17
Mean Average Precision.	17
R-precision.	17
2.2 Image Description	17
2.2.1 Saliency for image description	18
2.2.1.1 The Harris Corner Detector	19
2.2.1.2 Saliency from Local Complexity	19
2.2.1.3 Wavelet Based Saliency	20

2.2.1.4	Peaks in a difference-of-Gaussian Pyramid	21
2.2.1.5	Affine Covariant Region Detectors	22
	Harris-Affine and Hessian-Affine.	22
	Maximally Stable Extremal region detector (MSER).	24
	Affine Scale Saliency.	24
2.2.2	Image Features	24
2.2.2.1	Colour Features	25
2.2.2.2	Texture	25
2.2.2.3	Shape	26
2.2.2.4	Robust Local Descriptors - SIFT	26
2.3	The Semantic Gap and Auto-Annotation	27
2.3.1	Auto-Annotation Techniques	27
2.4	Summary	28
3	Image Description using Saliency	30
3.1	Requirements for Saliency Detectors for use in Robust Retrieval Scenarios	31
3.2	A Comparison of Saliency Detectors	31
3.2.1	Kadir's Scale-Saliency algorithm and Lowe's DoG-Peaks	31
3.2.1.1	Repeatability	32
	Repeatability Criterion.	33
3.2.1.2	Repeatability Results	34
3.2.2	DoG-Peaks and the State-of-the-Art Affine-Invariant Detectors	34
3.2.2.1	Image Data-set	37
3.2.2.2	Region Overlap and Repeatability	37
	Repeatability measure.	38
3.2.2.3	Matching	41
	Matching score.	41
3.2.2.4	Discussion of results	41
	General observations.	42
	Analysis of each transform.	43
	Conclusions.	48
3.3	A Simple Local Colour Descriptor	49
3.4	Summary	51
4	Image Retrieval using Salient Region Descriptors	52
4.1	Basic Model	52
4.1.1	Semantic Relevance	53
4.1.2	Results	55
4.1.3	Discussion	57
4.2	Text Retrieval Approaches	57
4.2.1	Applying Text Retrieval Techniques to Image Retrieval	57
4.2.1.1	Building visual words: Vector Quantisation	57
4.2.1.2	Image Retrieval based on visual words	60
	The Classical Approach	60
	The Latent Semantic Indexing Approach	61
4.3	Evaluation Techniques	61
4.3.1	Data-sets	61

4.3.2	Precision, Recall and Semantic Relevance	62
4.4	Results and Discussion	62
4.4.1	The Vocabulary	62
4.4.1.1	Vocabulary size	62
4.4.1.2	Sensitivity of retrieval with different vocabularies	64
4.4.2	Optimal k	65
4.4.3	Retrieval performance with the Washington data-set	65
4.4.4	Retrieval Performance with the Corel Data-set	69
4.4.5	Computational Performance	73
4.5	Conclusions	73
4.6	Summary	74
5	Query By Mobile Device	75
5.1	Requirements	76
5.2	Approach	78
5.2.1	Geometry-based Re-Ranking	78
5.2.2	Summary	79
5.3	Client-Server Implementation and Technology	79
5.4	Retrieval Performance	80
5.4.1	Discussion	84
5.5	Summary	84
6	Auto-Annotation and Advanced Retrieval	85
6.1	Auto-annotation using Semantic Propagation	86
6.1.1	Preliminary Results	86
6.1.1.1	Image Dataset	86
6.1.1.2	Performance Evaluation	87
6.1.1.3	Experimental Results	88
6.2	Using linear-algebra to associate images and terms	90
6.2.1	Decomposing the Observation Matrix	91
6.2.1.1	Interpreting the decomposition	93
6.2.2	Using the terms as a basis for new documents	93
6.2.3	Summary	94
6.2.4	A Simple Example	94
6.2.5	Some real examples	96
6.2.5.1	Building a training observation matrix	96
6.2.5.2	Experiments with the Washington data-set and SIFT ‘visual’ terms	98
	Choosing a good value for k	98
	Overall Retrieval Effectiveness.	98
	Example: Querying for “Bridge”.	100
6.2.5.3	The effect of including colour features in the Washington data-set	103
6.2.5.4	The Corel data-set	104
6.2.6	Discussion	110
6.3	Summary	112

7	Conclusions	113
7.1	Summary and Conclusions	113
7.1.1	Novel work in this Thesis	115
7.2	Future Work	116
7.2.1	Image Description using Saliency	116
7.2.2	Image Retrieval using Salient Region Descriptors	116
7.2.3	Query by Mobile Device	118
7.2.4	Auto-Annotation and Advanced Retrieval	118
7.3	The Future of CBIR	119
	Glossary	121
	Bibliography	124

List of Figures

2.1	An illustration of the Vector-Space model	7
2.2	An illustration of Latent Semantic Indexing	9
2.3	Graphical representation of dimensionality reduction	11
2.4	Examples of saliency operators	23
3.1	Comparison of difference-of-Gaussian and entropy response functions to a 1D signal	32
3.2	Effect of noise to entropy and difference-of-Gaussian response functions .	33
3.3	Sample image from the Washington data-set showing varying transforms .	35
3.4	Repeatability versus rotation and scale	36
3.5	Images of the Affine data-set showing viewpoint change	38
3.6	Images of the Affine data-set showing zoom and rotation	39
3.7	Images of the Affine data-set showing increasing image blurring	40
3.8	Images of the Affine data-set showing increasing JPEG compression . . .	40
3.9	Images of the Affine data-set showing illumination change	41
3.10	Detector results for viewpoint change with the Graffiti sequence	42
3.11	Detector results for viewpoint change with the Wall sequence	43
3.12	Detector results for scale change with the Boat sequence	44
3.13	Detector results for scale change with the Bark sequence	45
3.14	Detector results for blur change with the Bikes sequence	46
3.15	Detector results for blur change with the Trees sequence	47
3.16	Detector results for increasing JPEG compression with the UBC sequence	48
3.17	Detector results for illumination change with the Leuven sequence	49
3.18	The dominant colour descriptor applied to DoG regions on the first image of the Graffiti sequence	50
4.1	Sample images and their annotations from the Washington Ground Truth Image Database	54
4.2	Example showing retrieval with the DoG and global methods	56
4.3	Illustration of how the hue and saturation are quantised to form a vocab- ulary of colour ‘visual’ terms	59
4.4	Rank-frequency plot for ‘visual’ words	59
4.5	Generating vectors of occurrences of ‘visual’ terms from an image.	60
4.6	Precision-recall curves for different sizes of vocabulary using SIFT ‘visual’ terms with vector-space retrieval.	63
4.7	Precision-recall curves using the Washington data-set with vector-space retrieval	64

4.8	Relative R-Precision histograms showing the relative performance of retrieval using different vocabularies.	66
4.9	Effect of varying k with respect to retrieval performance for LSI based retrieval.	67
4.10	Precision-Recall for the Washington data-set with <i>SIFT</i> ‘visual’ terms . .	68
4.11	Precision-Recall for the Washington data-set with <i>colour</i> ‘visual’ terms . .	70
4.12	Precision-Recall for the Washington data-set with <i>combined</i> ‘visual’ terms	71
4.13	Precision-Recall for the Corel data-set	72
5.1	Screen-shot from the software demonstrator in capture mode	77
5.2	Montage showing various parts of the metadata shown to a user	78
5.3	Overview of our content-based image retrieval technique.	80
5.4	An overview of the mobile image retrieval system.	81
5.5	The system in use in a mock art gallery scenario	81
5.6	Example query images captured by the mobile device for testing the performance	82
5.7	Plot of the rank of the matching image for a number of different retrieval algorithms.	83
5.8	Retrieval rate versus N , the number of images considered for second-stage geometry based re-ranking.	83
6.1	Plot of empirical keyword distribution in the dataset	87
6.2	Precision-Recall curves for each of the auto-annotation methods	88
6.3	Example Annotations	89
6.4	Generating cross-language vectors of occurrences of ‘visual’ and annotation terms	97
6.5	The effect of k on average precision for four different queries.	99
6.6	The effect of k on the Mean-Average Precision over all 170 queries.	99
6.7	Average precision-recall curves for the different algorithms over all queries.	100
6.8	Average precision of Washington keyword queries	101
6.9	Relative R-Precision histograms of Factorisation versus Vector-Space . . .	102
6.10	Training images containing the “Bridge” keyword.	103
6.11	Precision-Recall curves for querying with the keyword “Bridge”	104
6.12	Test Images and their retrieved rank-order	105
6.13	Average precision-recall with different ‘visual’ terms	106
6.14	Relative R-Precision histogram between combined- and SIFT-‘visual’ terms	106
6.15	Plot illustrating the effect of varying k with the Corel data-set	107
6.16	Average Precision-Recall plots for the Corel data-set	108
6.17	R-Precision histograms for the Corel data-set	109
6.18	Precision-Recall curves for the top seven Corel queries	110

List of Tables

3.1	Number of regions detected by each detector for top-left image in Figure 3.5(a)	44
4.1	Averaged Semantic Relevance for queries based on the rank-1 result image and the closest 5 result images	55
4.2	Semantic Relevance for different sizes of vocabulary using SIFT ‘visual’ terms with vector-space retrieval.	63
4.3	Summary of average semantic relevance values for retrieval with the Washington data-set	67
6.1	Summary of Results	89
6.2	Comparison of precision between Factorisation and Machine Translation .	108

Nomenclature

N	Number of documents in a given corpus
\mathbf{V}_d	Vector representing document d in a vector-space
\mathbf{q}	Vector representing a query in a vector-space
\mathbf{A}	The term-document matrix
\mathbf{U}	Left-hand subspace from an SVD of \mathbf{A}
$\mathbf{\Sigma}$	Diagonal matrix of singular values from an SVD of \mathbf{A}
\mathbf{V}^T	Right-hand subspace from an SVD of \mathbf{A}
k	number of the space formed when performing Latent Semantic Indexing
\mathbf{U}_k	Reduced left-hand subspace from an SVD of \mathbf{A} constructed by considering only the first k -th columns
$\mathbf{\Sigma}_k$	Diagonal matrix of the largest k singular values from an SVD of \mathbf{A}
\mathbf{V}_k^T	Reduced right-hand subspace from an SVD of \mathbf{A} constructed by considering only the first k -th rows
\mathbf{A}_k	Rank k estimate of a term-document matrix formed calculating the product $\mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$
$\hat{\mathbf{q}}$	k -dimensional query vector formed by projecting a vector-space query vector, \mathbf{q} into a subspace, $\mathbf{q}^T \mathbf{U}_k \mathbf{\Sigma}_k^{-1}$
MAP	Mean Average Precision
$RP_A(i)$	R-Precision of the i -th query in system A
$RP_{A/B}(i)$	Relative R-Precision of the i -th query between system A and system B
$I(\mathbf{x})$	Intensity of the image I at the point \mathbf{x}
$I_x(\mathbf{x})$	Gradient (1st derivative) of the image I along the x -axis at the point \mathbf{x}
$I_{xx}(\mathbf{x})$	Gradient of the Gradient (2nd derivative) of the image I along the x -axis at the point \mathbf{x}
$I_{xy}(\mathbf{x})$	Derivative of the gradient image I_x along the y -axis at the point \mathbf{x}
$*$	Convolution operator
\mathbf{M}	Second moment matrix

H	Hessian matrix
σ	Standard deviation (of a Gaussian)
c	Constant factor relating to the ratio of σ between two Gaussian distributions in a difference-of-Gaussian
$D(\mathbf{x}, \sigma)$	Result of convolution of an image $I_x(\mathbf{x})$ with difference of Gaussian with standard deviations $(\sigma, c\sigma)$
H	Planar homography relating two images
X	Point in 3-Space
P_1, P_2	Projection matrices
p_1, p_2	Points in 2-space (on the image plane)
$ D(\epsilon) $	Cardinality of (number of elements in) the set $D(\epsilon)$
$r(\epsilon)$	repeatability of the points in $D(\epsilon)$
$D_E(\mathbf{F}_1, \mathbf{F}_2)$	Euclidean distance between feature vectors \mathbf{F}_1 and \mathbf{F}_2
$D_{\text{salient}}(\{\mathbf{F}_1\}, \{\mathbf{F}_2\})$	Distance between two equally sized sets of features $\{\mathbf{F}_1\}$ and $\{\mathbf{F}_2\}$
R_{semantic}	Semantic relevance
$V_{n,Z}$	Binary relevance of an image created by thresholding the semantic relevance above and below Z
E_{NS}	Normalised score measure
O	Observation matrix (analogous to a term-document matrix)
T	Term matrix, representing the locations of terms in a semantic space
D	Document matrix, representing the locations of documents in a semantic space
\mathcal{I}	Identity matrix
O*	Ideal, noise-free observation matrix
$\hat{\mathbf{T}}$	Estimated noise-free term matrix
$\hat{\mathbf{D}}$	Estimated noise-free document matrix
P	Partially observed observation matrix

Acknowledgements

Firstly, I would like to thank Paul Lewis for his support and supervision, without which this thesis would not have been possible. Paul's ideas and discussion have helped shape the work presented here.

Secondly I would like to thank the EPSRC and Motorola UK Research Laboratory for their continued support in funding this work. In particular, special thanks go to Paola Hobson, Angus Reid, Simon Waddington and Tony May of Motorola UK Research Laboratory.

I would also like to thank the members of the lab; in particular, Maria Karam, Simon Goodall, Christopher Bailey, Richard Lawley, Patrick Sinclair, Jiayu Tang, Wasara Rodhetbhai and Mark Thompson. Outside of the lab, I would particularly like to thank Jendra Gosai, Amanda Hatton, Deborah Santa Clara and Derya Sahin for their encouragement, suggestions and discussion.

In addition, I would like to thank the National Gallery for providing the data-set used of experimentation in Chapter 5 of this thesis, and graciously allowing us to use a number of the images as illustrations in this thesis and our other publications.

Finally I would like to thank my parents for their continued love and support.

Chapter 1

Introduction

“If I have seen further than others, it is by standing upon the shoulders of giants.”

SIR ISAAC NEWTON

The White Rabbit put on his spectacles. “Where shall I begin, please your Majesty?” he asked. “Begin at the beginning,” the King said, very gravely, “and go on till you come to the end; then stop.”

LEWIS CARROLL, ALICE’S ADVENTURES IN WONDERLAND

We live in the midst of the information age. Information is everywhere, and current society is beginning to require us to document everything, creating more information. In order to avert information overload, we need to develop techniques to search all this information.

This thesis is concerned in particular with visual information in the form of images. Even today, it is not uncommon for owners of digital cameras to have many thousands of photos stored on their personal computers. Mobile phones abound, and almost all modern phones come with built-in cameras of increasingly higher resolution. Through the internet it is possible to view millions of pictures created by others.

On the whole, these images have very little useful external metadata with which they can be indexed and searched, and so there is an increasing need for techniques to search these large image collections based on their content. This thesis attempts to investigate some of the issues involved with content-based search of image collections.

1.1 Aims and Objectives

The original aims of this work were to investigate how salient regions could be used in query-by-example retrieval scenarios where the query image was of particularly poor quality, such as in the case where it had been captured by a camera on a mobile device, such as a cellphone. This objective falls into two intertwined parts; the development of a robust image description using saliency, and the development of a retrieval approach to use this description.

With this objective complete, the work has evolved to investigate more advanced retrieval techniques, in the form of approaches that allow us to attack, or attempt to bridge the semantic gap. In this thesis, this has fallen into two very different techniques, directly inspired from the first objective. In the first of these techniques, we attempt to bridge the semantic gap by auto-annotation, that is, applying keywords to un-annotated images. In the second approach, we develop a linear-algebraic technique that essentially allows us to model the gap as a semantic space in which keywords and visual features are associated.

1.2 Contributions

This thesis brings a number of clear contributions to a number of fields, but in particular to the field of image retrieval. These contributions are itemised in brief below.

- A comparison of the Scale Saliency algorithm with the difference-of-Gaussian approach to finding salient regions.
- An in depth comparison of the difference-of-Gaussian algorithm to a number of state-of-the-art affine-invariant salient region detectors.
- The development of approaches to indexing images using descriptors of salient regions, in particular, approaches using information retrieval techniques adopted from the text retrieval field.
- Development of new techniques for assessing image retrieval performance when doing query by image content tasks in association with annotated test image sets.
- The development of an approach to improving the ranking of retrieved objects based on the existence of a planar homography between salient regions.
- The development of a demonstrator system that uses the above techniques to enable ‘query by mobile device’.
- Development of an approach to auto-annotation based on propagation of semantics from *similar* images.

- Formalisation and extension of an approach for text retrieval known as Cross-Language Latent Semantic Indexing which enables semantic spaces representing the relationships between observations of keywords and image features to be created using techniques from linear algebra. The technique allows un-annotated documents to be projected into the semantic space, uncovering hitherto unknown relationships, and allows these (un-annotated) documents in the space to be searched by keyword.

The research has led to four refereed conference publications on varying subjects, and one refereed workshop paper; Hare and Lewis (2003) discussed the applications of the Scale-Saliency algorithm (Kadir, 2001) for image matching, tracking and recognition/retrieval. Hare and Lewis (2004) described an evaluation of the Scale-Saliency algorithm and difference-of-Gaussian peaks detector (Lowe, 1999, 2004), and proposed a method of using the salient regions for query by example (QBE) tasks. The paper also proposed a new method for assessing retrieval performance of QBE tasks with annotated image sets. Hare and Lewis (2005a) demonstrated the idea of *query by mobile device* within an art gallery scenario, using content-based retrieval approaches evolved from Hare and Lewis (2004) using a vector-space retrieval model. Hare and Lewis (2005b) discussed the retrieval techniques described in Hare and Lewis (2005a) with respect to a more traditional retrieval environment. In addition the work was extended to cover another indexing approach called Latent Semantic Indexing (LSI), and a comparison was performed. Finally, Hare and Lewis (2005c) proposed a simple method for auto-annotation by propagation of keywords from *similar* images. Image similarity was assessed using both the vector-space and LSI indexing techniques.

1.3 Thesis Structure

This thesis describes the work of the author in attempting to achieve the objectives outlined earlier in this chapter. The early chapters of the thesis attempt to document and describe existing research towards these goals. Chapters 3 through 6 describe the actual research undertaken by the author, and Chapter 7 presents the conclusions of this research together with some views of the author regarding directions for future research. The following list describes the structure and content of the thesis on a chapter by chapter basis.

Chapter 2 - Background. Introduction to the background behind content-based retrieval, computational saliency, auto-annotation and the semantic gap. Also discusses techniques for assessing performance of retrieval and auto-annotation.

Chapter 3 - Image Description using Saliency. Research into the performance of different saliency detectors under varying transforms, concentrating in particular

on the performance of the difference-of-Gaussian peaks detector. A simple local-colour descriptor is also introduced.

Chapter 4 - Image Retrieval using Salient Region Descriptors. Investigates a number of techniques for image retrieval using the salient regions discussed in the previous chapter. The chapter culminates with the discussion of image retrieval techniques inspired by models from the text retrieval community.

Chapter 5 - Query by Mobile Device. Description of an example system that demonstrates the use of the techniques from the previous two chapters for image retrieval on a mobile device within an art gallery scenario. The system allows image queries to be captured using a camera built into the device and sent to a server for processing. The server returns metadata, such as a web-page corresponding to the closest matching image in a database, which is then displayed to the user on the screen of the device.

Chapter 6 - Auto-Annotation and Advanced Retrieval. Research into two advanced image retrieval strategies that attempt to bridge the semantic gap. The first strategy describes a simple auto-annotator using the techniques from Chapters 3 and 4. The second technique builds in particular on the Latent Semantic Indexing approach described in Chapter 4 in order to construct a semantic space that can be used to search for un-annotated images by keyword.

Chapter 7 - Conclusions. The overall results and contributions of the research from the previous four chapters is discussed, with respect to the original aims and objectives presented earlier in this chapter. The chapter ends in a discussion of future work with respect to all of the previous four chapters, but in particular to the research described in the second part of Chapter 6.

Chapter 2

Background

“We operate with nothing but things which do not exist, with lines, planes, bodies, atoms, divisible time, divisible space — how should explanation even be possible when we first make everything into an image, into our own image!”

FRIEDRICH WILHELM NIETZSCHE

This thesis uses techniques from a number of vast and multi-faceted fields, covering everything from information retrieval, to cognitive psychology (in the form of saliency), to computer vision. Whilst it would be far beyond the scope of this chapter to review all of these fields in depth, the chapter attempts to describe the techniques and *prior art* used throughout the remainder of the thesis.

The chapter begins by reviewing techniques in content-based retrieval; firstly textual information retrieval, and then image retrieval. This is followed by a discussion of techniques for image description, in particular techniques using salient points or regions. Finally the chapter looks at techniques for auto-annotation as an attempt to bridge what has been described as the *semantic gap*. The *semantic gap* can be described as the gap between low-level image descriptions, and the high-level semantics that the images convey and in which users typically prefer to articulate their queries.

2.1 Content-based Retrieval

Content-based retrieval is a technique for retrieving documents from a store such that the contents of the retrieved documents satisfy a user-provided information need. Unlike database retrieval, where a query is well defined and returns a set of records that exactly match the required specifications, content-based retrieval attempts to find objects or documents that are most similar to a specific query. The content-based retrieval process

usually involves generating signatures from the content of each of the documents in the archive or corpus, and comparing these signatures to the signature of the query. Results are usually ranked in terms of how *similar* their signatures are to the query signature.

The next part of this section describes two techniques for content based retrieval of textual documents, where the document signatures are created from vectors of the number of times each word in a lexicon occurs within the document. The final part of the section describes a number of techniques for the content-based retrieval of images.

2.1.1 Textual Information Retrieval

Archaeological evidence has suggested that man first begun organising information for later retrieval and usage over 4000 years ago. Examples of this include tables of contents in books. As the numbers of books increased, it became necessary to build specialised data structures to ensure fast data access. An old and popular data structure is the *index*, which contains a collection of words or concepts, and pointers to the related information. Traditionally, indexes have been manually created as forms of categorisation hierarchy which allow books of similar content to be grouped together, thus allowing a primitive form of content-based retrieval. Even today, libraries still use categorisation hierarchies, such as the Dewey decimal system (Dewey, 1876).

From around the mid to late 1960's, corresponding with the beginnings of the information age, research began on automatic computational approaches to text indexing and retrieval. The research led to three classic classes of models in information retrieval; the set theoretic Boolean models, algebraic vector models, and probabilistic models. The original vector model and a modern extension called Latent Semantic Indexing form a basis of this thesis, and are described next.

2.1.1.1 Classical Text Retrieval: The Vector Space Model

The vector-space model was developed by Salton et al. (1975). Most classical text retrieval systems work in the same general way, by representing a document and query as a set of terms. In the vector-space model, these terms are represented as axes in a vector space, using weighted term frequency as the distance along the axis corresponding to that term. Figure 2.1 illustrates the main idea behind the Vector-Space model, and the standard steps involved with creating this model are discussed below.

Parsing and Stemming. Firstly, a document is parsed into a list of separate words, this is obviously an easy task in most languages as the words are separated by spaces. The words are then transformed by a process called stemming. The stemming process represents words by their stems, for example, CONNECT, CONNECTED, CONNECTING,

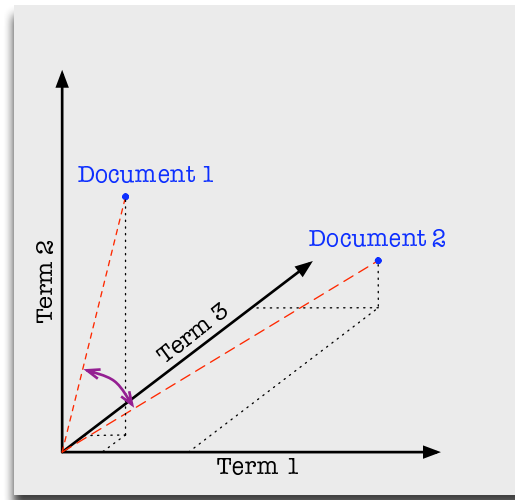


FIGURE 2.1: An illustration of the Vector-Space model; Document similarity is measured by angle between document vectors.

CONNECTION, AND CONNECTIONS are all represented by the stem CONNECT. Words with a common stem will often have similar meanings. Various algorithms for stemming have been developed, for example, the Porter Stemmer (Porter, 1980), which stems English words.

Stop Lists. The next stage is to apply a stop list. The stop list is used to reject common words which occur frequently throughout the corpus of documents, and therefore are not discriminating for a particular document. Examples of such words include words like ‘and’, ‘an’ and ‘the’.

Representing documents by word frequency. Each of the words from the document (after application of the stop list) are then represented by a unique identifier for that word. The number of occurrences of each word in the document is then counted and a vector of word-frequencies is created to represent the document.

Frequency weighting. Each component of the vector of word frequencies is often weighted. In the case of the Google web search engine, the weighting of terms within a particular web page depends on the position or class of the word within the page; for example, words in the title may be given a higher weight (Page and Brin, 1998).

A standard way of weighting the frequency vectors of text documents in the vector-space model is called ‘term frequency-inverse document frequency’, *tf-idf*, and is computed as follows¹. Suppose that there is a vocabulary of k words, then each document is represented by a k -dimensional vector $\mathbf{V}_d = (t_1, \dots, t_i, \dots, t_k)^T$ of weighted word frequencies

¹*tf-idf* actually refers to a class of different formulae for weighting terms, however, for simplicity we take it to mean the *basic* formulation as shown.

with components

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}, \quad (2.1)$$

where n_{id} is the number of occurrences of word i in document d , n_d is the total number of words in the document d , n_i is the number of documents in which the term i occurs in the whole database and N is the number of documents in the whole database. The weighting is the product of two terms: the *word frequency* n_{id}/n_d and the *inverse document frequency* $\log N/n_i$. The intuition is that word frequency increases the weights of words that occur frequently in a particular document, and thus describe it well, whilst the inverse document frequency down-weights words that appear often in the database.

Indexing using Inverted Files. Inverted file structures are used for efficient retrieval. An inverted file is like an ideal book index. Each word in the collection has an entry in the inverted file, together with a list of documents (and the positions in which the words occurs in them) that contain that word.

Searching: Ranking the results. In order to search the database of documents, a *tf-idf* vector \mathbf{q} is created for the query terms or document, and the query vector is compared against all the vectors \mathbf{V}_d in the database. The documents in the database are ranked using the normalised scalar product (cosine of angle):

$$\cos(\theta) = \frac{\mathbf{q} \bullet \mathbf{V}_d}{\|\mathbf{q}\| \|\mathbf{V}_d\|} \quad (2.2)$$

Term Rank - Term Frequency Plots and Zipf's Law. As mentioned previously, some words occur frequently — these are the words that tend to have little descriptive meaning and are often added to the stop list. Conversely, some words occur very infrequently in a document collection, but these words tend to be very descriptive of the content of the document. If, given a large corpus of documents written in some natural language (e.g. English), one were to count the frequencies of each word and plot a graph of rank frequency against frequency, one would find that the frequency of use of the n th-most-frequently used word is approximately inversely proportional to n . More specifically the graph will show a relationship of the form $f \propto 1/n^s$, where s is approximately one. This phenomenon is known as Zipf's Law, after linguist George Kingsley Zipf, who first observed the relationship. Plots of rank frequency versus frequency are useful in information retrieval as they help in choosing which words should occur in the stop list.

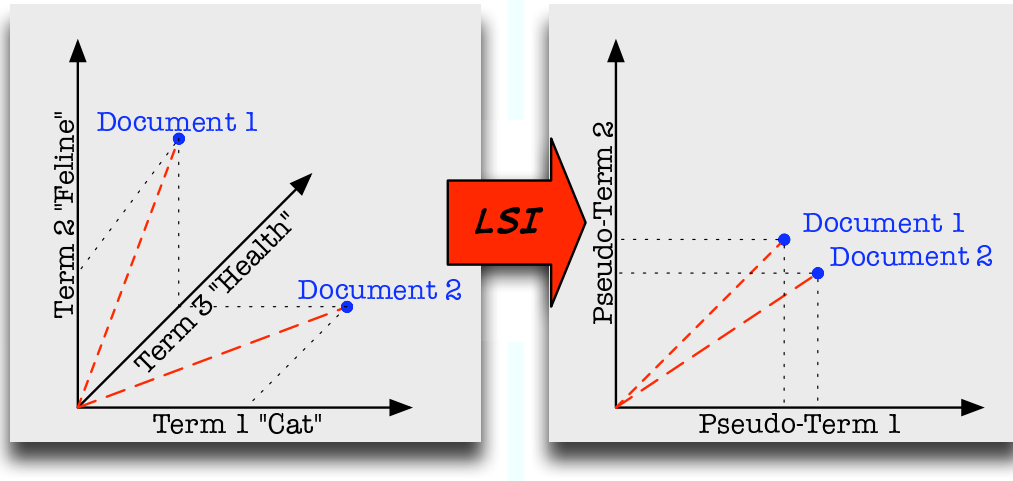


FIGURE 2.2: An illustration of Latent Semantic Indexing; LSI reduces the dimensionality so that similar documents have a smaller angle between their vectors.

2.1.1.2 The Vector Space Model Extended: Latent Semantic Indexing

The classical approach to text retrieval described above depends on a lexical match between the words in the query and those in the document collection. However, there is often a lot of diversity in the words used to describe a document (*synonymy*), and the words often have multiple meaning (*polysemy*), making the lexical methods incomplete and imprecise. Deerwester et al. (1990) suggest that it is possible to take advantage of the implicit higher-order structure in the association of terms with documents by determining the singular value decomposition (SVD) of large sparse term-by-document matrices. Terms and documents represented by the k largest singular vectors are then matched against user queries. Deerwester calls this retrieval method Latent Semantic Indexing (LSI) because the k -subspace represents important associative relationships between terms and documents that are not evident in individual documents (Berry et al., 1994). Figure 2.2 illustrates this idea.

The Term-Document Matrix and its Decomposition. LSI begins by constructing a vector space representation for each document, representing each document by a vector of word frequencies, as described in the previous section. The vectors are then arranged into a matrix \mathbf{A} , which is known as the term-document matrix. An individual element in \mathbf{A} , a_{ij} represents the frequency of term i in document j . The matrix \mathbf{A} is usually very sparse because every word does not normally occur in each document. It is normal to apply weightings to each element of \mathbf{A} , such that:

$$a_{ij} = L(i, j) \times G(i) \quad (2.3)$$

where $L(i, j)$ represents the local weighting for term i in document j and $G(i)$ is the global weighting for term i .

Log-Entropy Weighting. The most commonly used weighting for LSI is the “Log-Entropy” weighting. The local weighting is the log of the term-frequency of an individual document, and the global weighting is related to the entropy of the term frequency over the entire collection. This weighting scheme ensures that a term whose appearance tends to be equally likely among the documents is given a low weight and a term whose appearance is concentrated in a few documents is given a higher weight. The equations for the weighting are as follows:

$$L(i, j) = \log(tf_{ij} + 1) \quad (2.4)$$

$$G(i) = 1 - \sum_{j=1}^N \frac{\frac{tf_{ij}}{gf_i} \log(\frac{tf_{ij}}{gf_i})}{\log N}, \quad (2.5)$$

where tf_{ij} is the frequency of term i in document j , gf_i is the total number of times term i occurs in the entire collection, and N is the total number of documents in the collection.

Decomposition into a subspace. Once the weighted term-document matrix has been created, it is decomposed using the singular value decomposition. Briefly, SVD is used to decompose matrix \mathbf{A} into the product of three separate matrices, \mathbf{U} , $\mathbf{\Sigma}$, \mathbf{V}^T :

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (2.6)$$

The monotonically decreasing (in value) diagonal elements of the matrix $\mathbf{\Sigma}$ are called the singular values of the matrix \mathbf{A} . These matrices represent the breakdown of the original relationships into linearly-independent vectors or *factor values*. By selecting the first (largest) k singular values of \mathbf{A} , it is possible to construct a rank- k approximation to \mathbf{A} via $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$. This is illustrated in Figure 2.3. A theorem by Eckart and Young (1936)(see also Golub and Reinsch, 1971) suggests that the \mathbf{A}_k constructed from the largest k singular values of \mathbf{A} is the closest rank- k approximation (in the least squares sense) to \mathbf{A} . In terms of LSI, \mathbf{A}_k is the closest k -dimensional approximation to the original term-document space represented by \mathbf{A} . By reducing the dimensionality of \mathbf{A} , much of the “noise” that causes poor retrieval performance is thought to be eliminated.

Queries and Subspace Projection. In order to perform queries in the reduced term-document space, query vectors need to be represented as vectors in the k -dimensional space and compared to each document. Given a query vector, \mathbf{q} , whose non-zero elements contain the weighted (using the same weighting as in the creation of the term-document matrix) term-frequency counts of the terms that appear in the query, then, the query

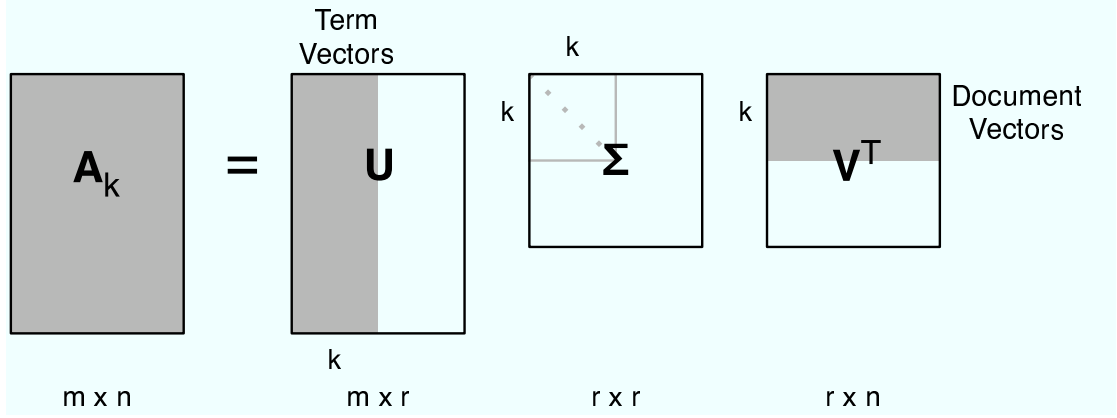


FIGURE 2.3: Graphical representation of dimensionality reduction of the term-document matrix using the singular value decomposition.

vector can be projected into the k -dimensional subspace:

$$\hat{\mathbf{q}} = \mathbf{q}^T \mathbf{U}_k \mathbf{\Sigma}_k^{-1} \quad (2.7)$$

The k -dimensional query vector, $\hat{\mathbf{q}}$ can then be compared against each of the document vectors and the results ranked. Again, a common similarity measure is the cosine similarity, described in Section 2.1.1.1.

2.1.2 Image Retrieval

To discuss the field of image retrieval in much detail would be far beyond the scope of this thesis. However, this section will attempt to give an overview of the salient facets and techniques of the field, drawing particular attention to the techniques built upon later in this thesis. Excellent reviews of image retrieval, and in particular content-based image retrieval can be found in the review articles by Smeulders et al. (2000) and Rui et al. (1999) and the report by Eakins and Graham (2000).

Active research into image retrieval has taken place since the late 1970's (Rui et al., 1999). Image retrieval has been approached from two main directions in the past by different communities; Database Management and Computer Vision. The Database Management community focused on techniques for retrieving images based on textual keywords or annotations, whilst the Computer Vision community investigated visual retrieval techniques. Nowadays, and into the future, these two areas of retrieval are becoming more intertwined, as discussed later in this chapter and in the concluding chapter of this thesis.

Many advances have been made over the years in the Database Management and textual information retrieval fields, such as the data modelling approaches discussed earlier and multidimensional indexing techniques. However, the use of annotations for image retrieval within this framework does suffer from two major difficulties, especially when

dealing with large numbers of images. The first of these problems is simply that of the expense of annotating each of the images. The second is related, and is about the subjectivity of the annotators; different people may perceive an image in different ways, and thus apply different annotations. This subjectivity may cause unrecoverable errors or mismatches in the retrieval process. The text-based image retrieval methodology is discussed in detail in the review papers by Chang and Hsu (1992) and Tamura and Yokoya (1984). The most pervasive text-based image search system available presently is perhaps the Google image search (<http://images.google.com>), which indexes images based on text and metadata from the web-page on which the image is embedded (Google Inc., 2005).

Over the last 15 years or so, the problems surrounding the text-based image retrieval approaches have become more and more acute due to the ever increasing size of image collections. The early 1990's saw the proposal of a new technique - Content-based Image Retrieval (CBIR). The aim of the content-based approach was to retrieve images relevant to a query, not by their keywords, but rather by their own visual content, such as the colours and textures within the image.

2.1.2.1 Applications of Image Retrieval

Smeulders et al. (2000) follows the categorisation of Cox et al. (2000) in describing the broad categories of user aims in image retrieval. Cox et al. (2000) describes these aims as *Target-Specific Search* or, simply, *Target Search*, *Category Search* and *Open-Ended Search - Browsing*.

In *Target Search*, users are required to find a specific image within a database; the search can only terminate when the specific image is found. Examples of where this type of search is valuable include checking whether a particular logo has been registered, searching for a particular photograph tied to a historical event, searching for a precise image in mind - as in searching art catalogues (e.g. Flickner et al., 1995), and, searching for a specific painting in order to find out the artist and title (e.g. Chan et al., 2001; Hare and Lewis, 2005a).

Category Search is where users search for images that belong to prototypical categories, such as “cities”, “sunsets” or “scenes of football games”. When a user is asked to find an image that is in someway *similar* to a target image they engage in a *category search*.

Browsing, or searching by association is where users search through a database with no particular specific goal in mind. Often, the goal of the search may change, and users may refine the search in an interactive, iterative manner though relevance feedback (Rui et al., 1997b, 1998; Hiroike et al., 1999). The search may start with specification by sketch (e.g. Kato et al., 1992) or by example image.

These three categories do not fully describe all of the aims of users when retrieving images, as shown by Armitage and Enser (1997). Enser (1995) attempts a more generic categorisation of image retrieval query requests from archives of still and moving imagery. Ornager (1997) studied how journalists retrieved images and identified five typical patterns. Ornager's patterns were classified as follows:

- The *specific inquirer* who asks very narrow questions, because he/she has a specific photograph in mind;
- The *general inquirer* who asks very broad questions because they want to make their own choice;
- The *story teller inquirer* who tells about the story and is open to suggestions from the archive staff;
- The *story giver inquirer* who hands the story over to the staff wanting them to choose the photograph(s); and
- The *fill in space inquirer* who only cares about the size of the photograph in order to fill an empty space on the page.

2.1.2.2 Retrieval Paradigms

The applications and user aims within content-based image retrieval described above have led to a number of paradigms or methods by which images can be retrieved. Some of these methods are listed below.

Browsing. Retrieval by browsing is perhaps the most commonly used paradigm. It is used by people often on a daily basis when trying to find information (not necessarily on a computer). The process is largely an iterative one in which the user gets closer to the information they require in an iterative manner by repeatedly selecting subsets of data. A common example of this is of a user searching for information in a library; The user will locate the appropriate section of the library (perhaps with the aid of a classification scheme, such as the Dewey Decimal System (Dewey, 1876)) in the first iteration, then select the appropriate shelves, then books, etc.

Navigation. Searching by navigation is restricted purely to the domain of computing with the advent of the *hyperlink*. This is best illustrated by the internet and web, where information can be sought by following links. The concept of the generic link proposed in the Microcosm system (Davis et al., 1992b,a) allows a link to provide a selection of documents to navigate to, where the documents are not hard-coded, but dynamically determined using the link anchor as a query passed to a retrieval engine. The MAVIS

(Microcosm Architecture for Video, Image and Sound) (Lewis et al., 1996b) provided generic linking for non-textual media; enabling linking based on the use of the media as the link anchor. The MAVIS II system (Lewis et al., 1996a; Dobie et al., 1999) extended the concept by incorporating a multimedia thesaurus, enabling navigation by concept.

Query By Example. Query by Example (QBE) is perhaps the most common form of retrieval in the content-based image retrieval community. The method allows users to specify queries in the form “find me documents *like* this one”. In addition to finding similarity matches, QBE can be used for finding exact matches (*Target Specific* searching).

Query By Sketch. Query by sketch allows the user to interactively generate a proxy document from which to perform a query by example style search. The proxy document generation could involve laying out shapes to indicate where particular colours should appear within the retrieved documents (Huang et al., 1996).

2.1.2.3 The Fundamental Bases of CBIR

Rui et al. (1999) describe three fundamental bases for Content-Based Image Retrieval. The bases are described briefly below.

Feature Extraction. The first stage of content-based image retrieval is to extract *features* from the image. These *features* represent some of the content of the image in some form. For example, the feature may describe the global colour distribution of the image. Feature extraction is described in more depth in Section 2.2.

High Dimensional Indexing. In order to make content-based image retrieval truly scalable, the extracted features have to be indexed in some manner. Proposed techniques for indexing have included tree structures, such as the R^* -tree (Beckmann et al., 1990) and priority k -d tree (White and Jain, 1996), clustering approaches (Charikar et al., 2004; Rui et al., 1997a), and neural network approaches (Zhang and Zhong, 1995). Some of the tree-based indexing techniques have been criticised in the literature because they break down when the number of dimensions exceeds about 20.

Image Retrieval System Design. The final base of content-based image retrieval is the construction of systems that combine the feature extraction and indexing stages in order facilitate searching using the retrieval paradigms discussed above. The CBIR review articles (Eakins and Graham, 2000; Rui et al., 1999; Smeulders et al., 2000) and the report by Venters and Cooper (2000) describe a number of image retrieval systems

in detail, including the first commercial system, QBIC (Flickner et al., 1995; Niblack et al., 1993; IBM Corporation, Accessed 10/9/2005), and the MARS system (Huang et al., 1996; Rui et al., 1997b).

2.1.2.4 Recognition and retrieval using salient interest points

Image description using saliency is described in detail in Section 2.2. However, the following descriptions give an overview of the *prior art* of the use of saliency in retrieval.

The ground-breaking work of Schmid and Mohr (1997) showed that it was possible to extend invariant local feature matching to general image recognition problems where a feature was matched against a large database of images. Schmid and Mohr used Harris corners (see Section 2.2.1.1) to detect interest points and used a local jet - a rotationally invariant feature descriptor to describe the characteristics of the local image region around the interest point. This demonstrated that it was possible to allow features to be matched under arbitrary orientation change between two images. It also showed that multiple feature matches could accomplish recognition under occlusion and clutter by identifying consistent clusters of matched points.

Harris corners are very sensitive to scale change, and so researchers began looking at other methods for selection of salient points. Lowe (2004) used peaks in the difference-of-Gaussian pyramid to select interest points and developed a highly distinctive local descriptor that is insensitive to small perturbations in location. He then went on to develop techniques for verifying object matches based on clusters of matching salient points. Shokoufandeh et al. (1999) developed a graph based matching and recognition strategy based on their wavelet based salient regions.

Sebe et al. (2003) introduce the idea of using salient points for content-based image retrieval. They used local features based on colour moments and Gabor texture features to describe the local characteristics around the salient interest points. The overall similarity measure between a query image and each database image was a linear combination of the similarity distance of each individual feature. Tuytelaars and Gool (1999) used locally affine invariant regions for their retrieval system. They rank returned images based on the number of votes an image received. Each point in the query image is matched to a point in one of the database images such that the Mahalanobis distance between the feature vectors is minimised. Each match between a point in the query image and a point in the database image is translated into a vote for that database image. Obdrzalek and Matas (2003) describe a retrieval system where affine invariant regions are computed and geometrically and photometrically normalised. These normalised regions are then described using low frequency components from a discrete cosine transform.

Each of the retrieval algorithms based on salient regions described above demonstrate a clear advantage over the use of global descriptors for image retrieval. Each of the methods showed a significant improvement in both retrieval accuracy and precision.

2.1.3 Retrieval Evaluation

Performance evaluation has become an increasingly important problem over the years. In the field of information retrieval, performance has often been measured by comparing how many documents returned for a query are actually relevant to that query. However, the problem with this is that the definition of what is relevant is subjective. To solve this problem, collections of documents must be created with distinct categories. This approach has been used with much success in the text retrieval community in conferences like TREC (Text REtrieval Conference), where there is a standard corpus of documents and categories, and a well defined protocol for retrieval engine evaluation. The TREC evaluation has motivated a similar effort for assessing content-based retrieval for video called TRECVID.

The most common measures of information retrieval are described below. Smith (1998) gives a review of these measures and more, with regards to content-based image retrieval.

2.1.3.1 Precision and Recall

The standard metrics for performance evaluation of information are called precision and recall. The precision of a query is defined as the ratio of the number of returned relevant documents to all documents returned by a retrieval system:

$$precision = \frac{|\text{retrieved relevant}|}{|\text{retrieved}|} \quad (2.8)$$

Recall is defined as the ratio of the number of retrieved relevant documents to the number of documents from the entire corpus that are relevant to the query:

$$recall = \frac{|\text{retrieved relevant}|}{|\text{relevant}|} \quad (2.9)$$

Precision and recall are related to the Receiver Operator Characteristic; Recall is the true positive rate, and precision is related to, but not the same as, the false positive rate. The precision and recall metrics have been applied to assessing the performance of content-based image retrieval systems. However, the metrics have a shortcoming in that the definition of the relevance of a document is assumed to be binary. This shortcoming is discussed in more detail in Chapter 4.

2.1.3.2 Single Value Summaries

Rather than comparing plots of precision and recall, it is sometimes useful to have a single value by which to compare the retrieval performance. The mean average precision (MAP) and R-Precision are two such values.

Mean Average Precision. The average precision is the average of the precision after each relevant document is retrieved:

$$\text{AveP} = \frac{\sum_{r=1}^N (P(r) \times \text{rel}(r))}{|\text{relevant}|}, \quad (2.10)$$

where r is the rank, $\text{rel}(r)$ is the binary relevance of the document with rank r , and $P(r)$ is the precision of that document. The Mean Average Precision is the average precision, AveP averaged over all queries.

R-precision. The R-precision is the precision after R documents have been retrieved, where R is the total number of documents relevant to the query. By definition, the recall at the R-precision is equal to the R-precision. The R-precision can be averaged over all queries, and in fact the averaged R-precision is highly correlated with the MAP (Aslam et al., 2005). The R-precision is useful for comparing two retrieval algorithms on a query-by-query basis. Let $RP_A(i)$ and $RP_B(i)$ be the R-precision values of two algorithms A and B for the i -th query. If we then define the difference, or relative R-precision, $RP_{A/B}(i)$ to be,

$$RP_{A/B}(i) = RP_A(i) - RP_B(i) . \quad (2.11)$$

Positive values of $RP_{A/B}(i)$ indicate that algorithm A has better performance for the query, negative values indicate algorithm B is better, and an $RP_{A/B}(i)$ equal to 0 indicates both algorithms perform equivalently. Multiple $RP_{A/B}(i)$ values for different queries can be plotted in the form of a histogram in order to give an overview of how the two algorithms perform relative to one another.

2.2 Image Description

Image description is the process of creating descriptions of the visual content of an image in a form that is useful to the problem being solved. In its lowest form, an image description, or signature, is a collection of one or more features that describe some aspect of the image content.

Much previous work in the field of content-based retrieval has been based around the concepts of using global descriptors to describe the content of the image. More recently

researchers have begun to realise that global descriptors are not necessarily good when it comes to describing the actual objects within the images and their associated semantics. Two approaches have grown from this realisation; firstly approaches have been developed whereby the image is segmented into multiple regions, and separate descriptors are built for each region; and secondly, the use of salient points has been suggested.

The first approach has been demonstrated to work (Carson et al., 2002), although it has a large problem — that of how to perform the segmentation. Over the years many techniques for performing image segmentation have been suggested, although none really solve the problem of linking the segmented region to the actual object that is being described. Indeed, this shows that the non-naive segmentation problem is not just a bottom-up image processing problem, but also a top-down problem that requires knowledge of the true object before it can be successfully segmented.

The second approach avoids the problem of segmentation altogether by choosing to describe the image and its contents in an altogether different way. By using salient points within an image, it is possible to derive a compact image description based around the local attributes of the salient points. A number of different methods for finding salient points have been suggested, from the simple Harris and Stephens (1988) corner detector, to wavelet based approaches (Shokoufandeh et al., 1999; Sebe et al., 2003; Sebe and Lew, 2003), to methods centred around image entropy (Kadir, 2001; Kadir and Brady, 2001). Many previous approaches to using salient points have generated feature-vectors from pixel data in fixed-sized regions around the salient point, usually a 3×3 or 9×9 pixel neighbourhood centred on the point (Sebe et al., 2003), although some of the modern state-of-the-art detectors find affine invariant regions and generate descriptors from within the region (Tuytelaars and Gool, 1999; Sivic and Zisserman, 2003; Obdrzalek and Matas, 2003).

2.2.1 Saliency for image description

There exist a number of pre-requisites for the performance of saliency detectors that can be used in the context of image retrieval and recognition. The main requirement is one of repeatability, that is the same salient interest points should be selected regardless of imaging conditions and transformations, such as those from a small change camera location. A full mathematical definition of repeatability can be found in Chapter 3 (Section 3.2.1.1) together with a discussion of some of the other requirements for saliency detectors. What follows is a description of a number of various different salient interest point and region detectors that are applicable to recognition and retrieval.

2.2.1.1 The Harris Corner Detector

The interest point detector developed by Harris and Stephens (1988) is perhaps the most widely cited and ubiquitous of all interest point detectors. It is often used as a baseline for comparing the performance of newer detectors. The Harris Corner detector works by considering the second moment, or auto-correlation, matrix:

$$\mathbf{M} = \mu(\mathbf{x}) = \begin{bmatrix} I_x^2(\mathbf{x}) & I_x I_y(\mathbf{x}) \\ I_x I_y(\mathbf{x}) & I_y^2(\mathbf{x}) \end{bmatrix} \quad (2.12)$$

where $I(\mathbf{x})$ is the grey level intensity of the image at point \mathbf{x} and $I_x(\mathbf{x})$ is the derivative of I in the direction of the x -axis at the point \mathbf{x} . Similarly, $I_y(\mathbf{x})$ is the derivative of I in the direction of the y -axis at the point \mathbf{x} . If at a certain point the two eigenvalues of the matrix \mathbf{M} are large, then a small motion in any direction will cause an important change in grey level. This indicates that the point is a corner. The corner response function is given by:

$$R = \det \mathbf{M} - k(\text{trace } \mathbf{M})^2 \quad (2.13)$$

where k is a parameter set to a value of 0.04 (a suggestion of Harris). Corners are defined as local maxima of the corner response function. Sub-pixel accuracy can be achieved through quadratic approximation of the local neighbourhood of the local maxima. Corners due to image noise can be avoided by smoothing the images containing the squared derivatives $(I_x^2(\mathbf{x}), I_y^2(\mathbf{x}), I_x I_y(\mathbf{x}))$ with a Gaussian filter. Often the corner response function finds too many corners, so the number of corners is often reduced by applying non-maximal suppression and/or only selecting R values above a certain threshold.

The performance of the Harris detector is limited by the ability to estimate the image derivatives in a robust and rotationally insensitive manner. Often, corners found when the image is horizontal will not be found if the image is rotated by 45° in the plane.

Figure 2.4(a) illustrates the results of applying the Harris detector to an image.

2.2.1.2 Saliency from Local Complexity

Gilles (1998) investigated salient local image patches or ‘icons’ to match and register two images (specifically aerial reconnaissance images). Gilles suggested that by extracting locally salient features from the pair of images and matching these, it would be possible to estimate the global transform between the two images. Gilles defined saliency in terms of local signal complexity or unpredictability. More specifically, Gilles suggested the use of Shannon Entropy of local attributes to estimate the saliency. Basically, image segments

with flatter intensity histogram distributions² tend to have higher signal complexity and thus higher entropy. Gilles' method only worked at a single scale, and picked single salient points, rather than salient regions.

Kadir and Brady (2001) (see also Kadir, 2001) modified Gilles original algorithm to make it perform well on images other than those from aerial reconnaissance imagery. Essentially they changed the algorithm so that it detected salient regions at multiple scales. The modified algorithm located circular patches of the original image that were considered salient. The size of the patch was determined automatically by the multi-scale additions to Gilles' algorithm. In addition Kadir and Brady developed a simple clustering algorithm to group together features within the \mathbb{R}^3 space that have similar x and y location, and scale.

In more detail, the scale-saliency algorithm works by considering circular regions \mathcal{R} of radius, or scale, s , centred at a point \mathbf{x} within the image $I(\mathbf{x})$. The entropy, \mathcal{H} , of each region is calculated from an estimate of the probability density function of pixel intensities, $p(I, s)$ over \mathcal{R} , as follows:

$$\mathcal{H} = - \sum_I p(I, s) \log(p(I, s)) \quad (2.14)$$

The set of extrema with respect to scale in \mathcal{H} is computed over a range of the s parameter for all pixels in the image. For each extremum, a weighting \mathcal{W} is calculated as

$$\mathcal{W} = s \sum_I |p(I, s) - p(I, s + 1)|. \quad (2.15)$$

The *saliency*, \mathcal{Y} , of each circular region is calculated as $\mathcal{Y} = \mathcal{H}\mathcal{W}$. Kadir's implementation then applies a simple clustering algorithm to cluster together regions with similar spatial location and scale. Figure 2.4(b) illustrates the results of applying the algorithm to an image.

2.2.1.3 Wavelet Based Saliency

Wiscott et al. (1997) used Gabor wavelet jets to extract salient features for their face recognition algorithm. Wavelet jets represent an image patch, containing a feature of interest, with a set of wavelets across the frequency spectrum. Each set of wavelet responses represents a node in a grid-like planar graph covering overlapping regions within the image, which is in itself a kind of saliency map.

Shokoufandeh et al. (1999) use dyadic multiscale wavelets to find the scale which captures the most efficient encoding of an object's salient shape. Essentially, a saliency map

²Kadir and Brady (2001) note that the method is not limited to the intensity histogram and that it is equally possible to use a histogram from a different descriptor, such as colour or edge strength.

is created for each dyadic scale based on the wavelet response and a function that defines whether that scale best encodes an object's shape. Shokoufandeh et al. (1999) demonstrate the method to find circular patches at each scale.

2.2.1.4 Peaks in a difference-of-Gaussian Pyramid

The idea of using peaks in a difference-of-Gaussian pyramid comes from the work of Lowe (2004, 1999) on object recognition using keypoints. Lowe has shown that by searching a difference-of-Gaussian pyramid for local peaks, both spatially and across scale, it is possible to select points robust to a range of projective transformations.

Koenderink (1984) and Lindeburg (1994) showed that under a variety of reasonable assumptions, the only possible scale-space kernel is a Gaussian function. Therefore, the scale-space of an image is a function $L(\mathbf{x}, \sigma)$, that is produced from the convolution of a variable scale Gaussian, $G(\mathbf{x}, \sigma)$ and the image $I(\mathbf{x})$,

$$L(\mathbf{x}, \sigma) = G(\mathbf{x}, \sigma) * I(\mathbf{x}), \quad (2.16)$$

where $*$ represents the convolution operation and the 2D Gaussian kernel is given by:

$$G(\mathbf{x}, \sigma) = G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad (2.17)$$

Lowe (1999) proposed that stable interest points (or, in fact regions) could be selected by locating scale-space peaks in the difference-of-Gaussian function convolved with the image, $D(\mathbf{x}, \sigma)$, which can be computed from the difference of two nearby scales separated by a constant factor c :

$$\begin{aligned} D(\mathbf{x}, \sigma) &= (G(\mathbf{x}, c\sigma) - G(\mathbf{x}, \sigma)) * I(\mathbf{x}) \\ &= L(\mathbf{x}, c\sigma) - L(\mathbf{x}, \sigma) \end{aligned} \quad (2.18)$$

The difference-of-Gaussian closely approximates the scale-normalised Laplacian-of-Gaussian, $\sigma^2 \nabla^2 G$ (Lindeburg, 1994; Marr, 1982; Lowe, 2004). Lindeburg (1994) showed that a normalisation of the Laplacian by a factor of σ^2 was required for true scale invariance. Mikolajczyk (2002) showed that the minima and maxima of $\sigma^2 \nabla^2 G$ produced the most stable interest points when compared to a range of other operators.

In order to select peaks in the scale space, Lowe suggested testing each sample point to find out if it was larger or smaller than all its eight closest neighbours in image location and nine neighbours in the scale above and below. Once the scale-space peaks have been selected, Lowe suggests that the peaks can be better localised by fitting a 3D quadratic function to the local neighbourhood of the peak and finding the maxima. Lowe also

suggests that poorly defined peaks in the difference-of-Gaussian scale space should be rejected.

A poorly defined peak will have a large principle curvature across the edge, but a small one perpendicular to it. The principle curvatures are proportional to the eigenvalues of the 2×2 Hessian matrix, \mathbf{H} computed at the location and scale of the interest point:

$$\mathbf{H} = \begin{bmatrix} D_{xx}(\mathbf{x}, \sigma) & D_{xy}(\mathbf{x}, \sigma) \\ D_{xy}(\mathbf{x}, \sigma) & D_{yy}(\mathbf{x}, \sigma) \end{bmatrix} \quad (2.19)$$

The eigenvalues need not be calculated explicitly, as only the ratio of the eigenvalues is important. It can be shown that in order to test that the ratio of principle curvatures is below some threshold, then this is equivalent to checking

$$\frac{\text{trace}(\mathbf{H})^2}{\det \mathbf{H}} < \frac{(r+1)^2}{r}, \quad (2.20)$$

where r is the ratio between the smallest and largest eigenvalues of \mathbf{H} . Lowe suggests setting $r = 10$ which eliminates interest points that have a ratio of principle curvature greater than 10.

It should be noted that whilst Lowe only refers to interest points, the selection of peaks from the difference-of-Gaussian pyramid actually selects regions of the image, where the size of the region is related to the scale of the interest point. Figure 2.4(c) illustrates the results of finding peaks in a difference-of-Gaussian pyramid.

2.2.1.5 Affine Covariant Region Detectors

A number of recent state-of-the-art techniques have been suggested that are able to detect regions that are invariant to affine transforms (Tuyltaars and Gool, 1999; Obdrzálek and Matas, 2002; Mikolajczyk, 2002). However, these approaches are not yet fully affine invariant as they start with initial feature scales and locations selected in a non-affine-invariant manner. Mikolajczyk (2002) showed that the performance of his affine invariant detector was below that of the difference-of-Gaussian peaks detection method, until the difference in viewpoint of the two images being matched was very large. A small section of affine-covariant region detectors is discussed here.

Harris-Affine and Hessian-Affine. Both the Harris-Affine and Hessian-Affine detectors work in a similar manner, by selecting initial points then selecting scale. An iterative approach then selects elliptical regions based on the eigenvalues of the second moment matrix (c.f. Equation 2.12). The iteration stops when the eigenvalues of the second moment matrix calculated from pixels of the elliptic region (normalised to a circle) are equal. The Harris-affine detector, as its name suggests, selects initial points

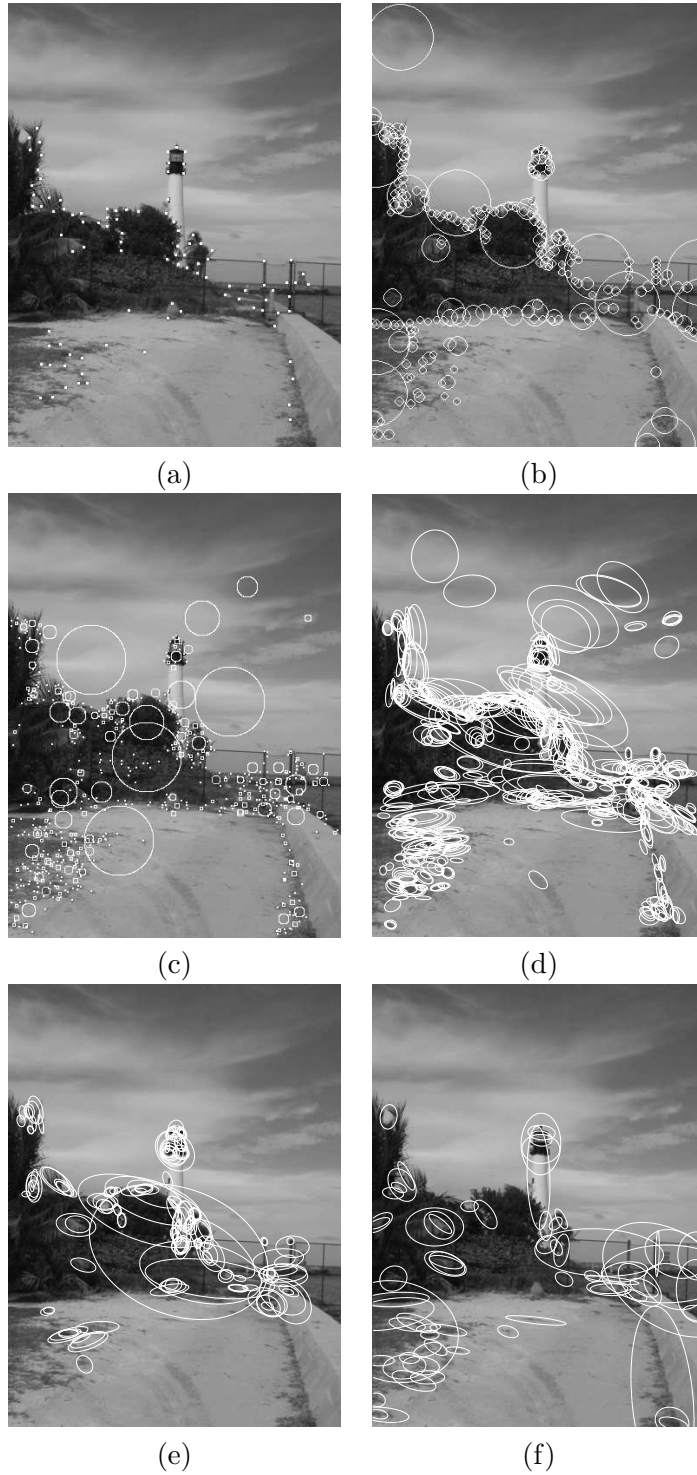


FIGURE 2.4: (a) Interest points found by the Harris Corner Detector; (b) Salient regions found by the Scale-Saliency algorithm; (c) Salient regions found by from peaks in a difference-of-Gaussian pyramid (region radius is equal to the size of the smaller σ in the difference-of-Gaussian). (d) Affine-covariant regions detected by the Hessian-Affine detector. (e) Affine-covariant regions detected by the Harris-Affine detector. (f) Elliptical Regions fitted to the regions detected by the MSER detector.

using the same technique as the Harris detector. The Hessian-affine detector selects points in a similar manner, but instead of selecting points based on the eigenvalues of the second moment matrix, points are selected based on the determinant of the Hessian matrix (c.f. Equation 2.19). The second derivatives used by the Hessian matrix give strong responses to ridge and blob structures, these are very similar to those detected by the Laplacian operator. The use of the determinant penalises very long ridge-like structures where the second derivative in one particular orientation is small.

The scale of the interest point is selected by choosing the characteristic scale at which the local structure gives the maximum response of a Laplacian operator. Figure 3.1 in Chapter 3 illustrates this idea by showing how the response of a difference-of-Gaussian operator (closely related to the Laplacian) to a simple one-dimensional signal over a range of scales. Figure 2.4(d) and (e) show regions found by the Hessian-Affine and Harris-Affine detectors within a sample image.

Maximally Stable Extremal region detector (MSER). The Maximally Stable Extremal Region detector was developed by Matas et al. (2002). The detector finds arbitrarily shaped regions in the form of connected components of an appropriately thresholded image. The regions are extremal because all of the surround pixels have either higher or lower intensity than the pixels within the region. The regions are *maximally stable* because of the optimal threshold selection process. The stability is measured as a function of how stable the local binarisation of the pixels is over a range of thresholds. As the threshold changes, the number of pixels within a connected region will likely change as well; if the number of pixels is fairly constant, then the region is stable. This definition of region stability based on relative area change is affine-invariant. Figure 2.4(f) shows elliptical regions fitted to the MSER regions in a sample image using the method described by Mikolajczyk et al. (2005).

Affine Scale Saliency. Kadir et al. (2004) presented an extension to the original Scale Saliency algorithm. The modifications involved changing the sampling region from a circle parameterised by its centre and radius (scale) to an ellipse parameterised by its scale (length of the major axis), orientation (of the major axis) and the ratio of major to minor axes. The original clustering algorithm was upgraded with an improved greedy algorithm.

2.2.2 Image Features

In order to create an image description, one has to extract features from the image. As discussed previously, features can be global, describing a characteristic of the entire image, or they can be local, describing a characteristic of a segmented- or salient- region.

There are also *pseudo*-global descriptors that describe the whole of an image, but are built from the specific arrangement of regions and their descriptors within the image. It is also possible to classify features as being general, or domain-specific. General features include things such as colour and texture, whilst the domain-specific features may describe such things as faces or fingerprints. From a retrieval standpoint, it is often better to combine multiple features to generate a more robust image description. Some common image features used in content-based image retrieval are described below.

2.2.2.1 Colour Features

Colour is perhaps the most widely used of all visual features in image retrieval. Most colour feature representations are relatively robust to image size and orientation. Colour is most often indexed in the RGB or HSV colour-spaces, however other *perceptual* colour-spaces have also been suggested. Finlayson et al. (1998) discuss colour-normalisation techniques for indexing.

By far the most common colour descriptor (used both globally and locally) is the colour histogram first proposed for use in retrieval by Swain and Ballard (1991). Stricker and Orengo (1995) noted most colour histograms are sparse and sensitive to noise, and suggested using the cumulative colour histogram instead, which they showed to be insensitive to the quantisation parameter. Stricker and Orengo also proposed a second technique in which only the dominant features of the colour distribution were indexed, in the form of *colour moments* from the first three moments (mean, variance and skewness) of the colour histogram. Sebe et al. (2003) used local colour moment descriptors together with salient points for retrieval.

Smith and Chang (1995) proposed the Colour Set feature formed from a set of colours from a quantised colour-space. The Colour Set features were binary, and thus allowed a binary search tree to be constructed for fast search (Smith and Chang, 1996).

Pass et al. (1996) take a two stage approach to indexing in which the image is segmented by reducing the number of colours. Pixel values of segmented regions with large areas are then stored in a *coherent* vector, and those from small regions are stored in a *incoherent* vector. Results showed this approach worked better than the simple colour histogram.

2.2.2.2 Texture

Texture in an image refers to homogeneous visual patterns within the image that are not due to a single colour or intensity. Haralick et al. (1973) was perhaps the first to suggest the use of texture as a feature, with the co-occurrence representation that explored the spatial relationships between grey-level pixels. Tamura et al. (1978) investigated computational approximations of texture properties found to be important from psychological

studies. These Tamura textures are attractive for image retrieval because they are visually meaningful. The Tamura textures were exploited in both the MARS (Huang et al., 1996) and QBIC (Niblack et al., 1993) retrieval systems. Howarth and Rüger (2004) carried out a detailed evaluation of the use of textures in a query-by-example image retrieval task.

Textures have also been represented using the Wavelet transform (e.g. Smith and Chang, 1994; Laine and Fan, 1993). In particular, Ma and Manjunath (1995) showed that the Gabor Wavelet transform performed well in a texture annotation task.

2.2.2.3 Shape

Shape is important in some retrieval scenarios, such as trademark retrieval (Eakins et al., 1998). Eakins (1993) discusses some design requirements for a shape retrieval system. Shape-based retrieval does suffer from the drawback that it requires an initial segmentation to select the shapes from the image.

In general, shape descriptors can be separated into two categories; region-based and boundary-based. Perhaps the most successful region-based descriptors are moment invariants introduced by Hu (1962). The characteristic boundary-based descriptor is the Fourier descriptor (Zahn and Roskies, 1972).

2.2.2.4 Robust Local Descriptors - SIFT

There are a large number of different types of feature descriptors that have been suggested for describing the local image content within a salient region; For example colour moments and Gabor texture descriptors (Sebe et al., 2003; Stricker and Orengo, 1995; Ma and Manjunath, 1995). However, many of these descriptors are not robust to poor imaging conditions. A study by Mikolajczyk and Schmid (2003) showed that the Scale Invariant Feature Transform (SIFT) descriptor, designed by Lowe (2004), was superior to other descriptors found in the literature, such as the response of steerable filters or orthogonal filters. The performance of the SIFT descriptor is enhanced because it was designed to be invariant to small shifts in the position of the sampling region, as might happen in the presence of imaging noise.

The SIFT descriptor is a three-dimensional histogram of gradient location and orientation. Lowe, suggests that gradient location be quantised into a 4×4 location grid, and gradient angle be quantised into 8 orientation bins. The resulting descriptor has 128 dimensions. Illumination invariance is obtained by normalising the descriptor by the square root of the sum of the squared components.

2.3 The Semantic Gap and Auto-Annotation

The hallmark of a good retrieval system is its ability to respond to a user's queries and present results in a desired fashion. In the past there has been a tendency for research to focus on content-based retrieval techniques, ignoring the issues of users. In spite of this, some investigators have attempted to characterise image queries, providing insights in retrieval system design (Enser, 1995; Armitage and Enser, 1997; Ornager, 1997; Hollink et al., 2004) and highlighting the problem of what has become known as the *semantic gap*.

In the survey of content-based image retrieval by Smeulders et al. (2000), the semantic gap is described as;

...the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.

At the end of the survey the authors conclude that:

A critical point in the advancement of content-based retrieval is the semantic gap, where the meaning of an image is rarely self-evident. The aim of content-based retrieval systems must be to provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics.

Techniques for attempting to bridge the semantic gap in image retrieval have mostly used an *auto-annotation* approach, in which keyword annotations are applied to unlabelled images. Enser et al. (2005) discusses some short-comings of auto-annotation due to their lack of *richness* when compared to real image annotations in archival collections. Enser et al. (2005) goes on to suggest that perhaps a way forward is to combine shareable ontologies to make explicit the relationships between the keyword labels and concepts they represent (e.g. Addis et al., 2003; Goodall et al., 2004; Hu et al., 2003). Zhao and Grosky (2000) proposed an approach to bridging the semantic gap using Latent Semantic Indexing (see also Grosky and Zhao, 2001; Cascia et al., 1998) — an approach that is further explored in this thesis.

2.3.1 Auto-Annotation Techniques

The first attempt at automatic annotation was perhaps the work of Mori et al. (1999), which applied a co-occurrence model to keywords and low-level features of rectangular

image regions. The current techniques for auto-annotation generally fall into two categories; those that first segment images into regions, or ‘blobs’ and those that take a more scene-orientated approach, using global information. The segmentation approach has recently been pursued by a number of researchers. Duygulu et al. (2002) proposed a method by which a machine translation model was applied to translate between keyword annotations and a discrete vocabulary of clustered ‘blobs’. The data-set proposed by Duygulu et al. (2002) has become a popular benchmark of annotation systems in the literature. Jeon et al. (2003) improved on the results of Duygulu et al. (2002) by recasting the problem as cross-lingual information retrieval and applying the Cross-Media Relevance Model (CMRM) to the annotation task. Jeon et al. (2003) also showed that better (ranked) retrieval results could be obtained by using probabilistic annotation, rather than *hard* annotation. Lavrenko et al. (2004) used the Continuous-space Relevance Model (CRM) to build continuous probability density functions to describe the process of generating blob features. The CRM model was shown to outperform the CMRM model significantly. Metzler and Manmatha (2004) propose an inference network approach to link regions and their annotations; unseen images can be annotated by propagating belief through the network to the nodes representing keywords. The models by Monay and Gatica-Perez (2003), Feng et al. (2004) and Jeon and Manmatha (2004) use rectangular regions rather than blobs. Monay and Gatica-Perez (2003) investigates Latent Space models of annotation using Latent Semantic Analysis and Probabilistic Latent Semantic Analysis, Feng et al. (2004) use a multiple Bernoulli distribution to model the relationship between the blocks and keywords, whilst Jeon and Manmatha (2004) use a machine translation approach based on Maximum Entropy. Blei and Jordan (2003) describe an extension to Latent Dirichlet Allocation (Blei et al., 2003) which assumes a mixture of latent factors is used to generate keywords and blob features. This approach is extended to multi-modal data in the article by Barnard et al. (2003).

Oliva and Torralba (2001) and Oliva and Torralba (2002) explored a scene oriented approach to annotation in which they showed that basic scene annotations, such as ‘buildings’ and ‘street’ could be applied using relevant low-level global filters. Yavlinsky et al. (2005) explored the possibility of using simple global features together with robust non-parametric density estimation using the technique of kernel smoothing. The results shown by Yavlinsky et al. (2005) were comparable with the inference network (Metzler and Manmatha, 2004) and CRM (Lavrenko et al., 2004). Notably, Yavlinsky et al. showed that the Corel data-set proposed by Duygulu et al. (2002) could be annotated remarkably well by just using global colour information.

2.4 Summary

This chapter has introduced a wide range of ideas and techniques, all broadly related to the field of content-based image retrieval. Information retrieval techniques, including the

vector-space model and Latent Semantic Indexing were introduced. This was followed by a discussion of techniques for image retrieval. The topic of image description was described, with particular emphasis on the use of saliency for robust image description. Finally, the chapter concluded with a discussion of the *semantic gap* in image retrieval, and some of the automatic annotation techniques that have been developed to attempt to at least partially bridge the gap.

Chapter 3

Image Description using Saliency

“One picture is worth ten thousand words.”

FREDERICK R. BARNARD

The use of saliency in computer vision has become quite widespread in recent years. Saliency is often used to provide the basis for a visual attention mechanism that reduces the need for computational resources. Historically, saliency was described by the term ‘interest point detectors’, but the use of the term ‘saliency’ has come about from the large amount of psychology-based work on selective visual attention.

Primates appear to solve much of the problem of visual scene analysis and object recognition in a serial manner. This approach is slower, but less computationally intensive than a parallel approach (Salah et al., 2002). This process is often referred to as *selective visual attention*. The idea of selective visual attention is that not all parts of an image give equal amounts of information, and that analysing only the relevant parts in detail is sufficient for recognition, retrieval and analysis.

As mentioned in Section 2.2.1, the use of saliency for image description has certain advantages over other approaches, such as global and segmentation-based description.

This chapter starts by discussing some of the requirements for saliency detectors in the context of image retrieval and goes on to compare the performance of a number of saliency/interest-point detectors. The third section of this chapter discusses a simple local colour descriptor developed for the retrieval scenarios presented in the later chapters of the thesis. Finally the chapter ends with a brief summary of the salient features and findings.

3.1 Requirements for Saliency Detectors for use in Robust Retrieval Scenarios

In a given content-based image retrieval scenario the aim is to find an image or images that are in some sense similar to a query image. If the images are represented by salient regions, then the aim is to find images with *similar* salient regions. The definition of *similar* in this case can have two different meanings; we can look for similar spatial arrangements of salient regions, or we can look for images with similar content, based on descriptors of the regions. Sometimes the retrieval scenario may even call for these two definitions to be combined, as demonstrated in Chapter 5.

Both of these cases have similar requirements of the region detectors. First and foremost, regions must be repeatable. Given an image and a geometrically- and/or photometrically-transformed version of it, regions detected in the first image should be detected in corresponding locations in the second image. The kinds of transformations to be expected are a function of the retrieval scenario, but typical transformations include the addition of noise, change in viewpoint, rotation, scaling, blurring, illumination changes, and compression. When retrieval is to be performed based on local content descriptors of salient regions, the local descriptors also need to be robust to the transformations typical of the retrieval scenario.

Another important factor for some retrieval scenarios is that of the distinctiveness of the regions with respect to the descriptors used to describe them. Take for example the use of a histogram of pixel values as the local descriptor. A salient region detector that picks regions with largely homogeneous content is unlikely to give very distinct descriptors, whereas a detector that picks regions with variable content will be much more distinctive.

3.2 A Comparison of Saliency Detectors

This section describes the results of two in-depth comparisons between a number of different saliency detectors. In the first subsection, a comparison of Kadir and Brady's Salient Scales algorithm and Lowe's difference-of-Gaussian Peaks method is described (Hare and Lewis, 2004). In the second subsection, Lowe's difference-of-Gaussian Peaks method is compared to six state-of-the-art affine-invariant region detectors using the methodology and data-set proposed by Mikolajczyk et al. (2005).

3.2.1 Kadir's Scale-Saliency algorithm and Lowe's DoG-Peaks

Both Kadir's and Lowe's methods for selecting salient regions are conceptually quite similar because they respond to a signal in the same way. For example, when the

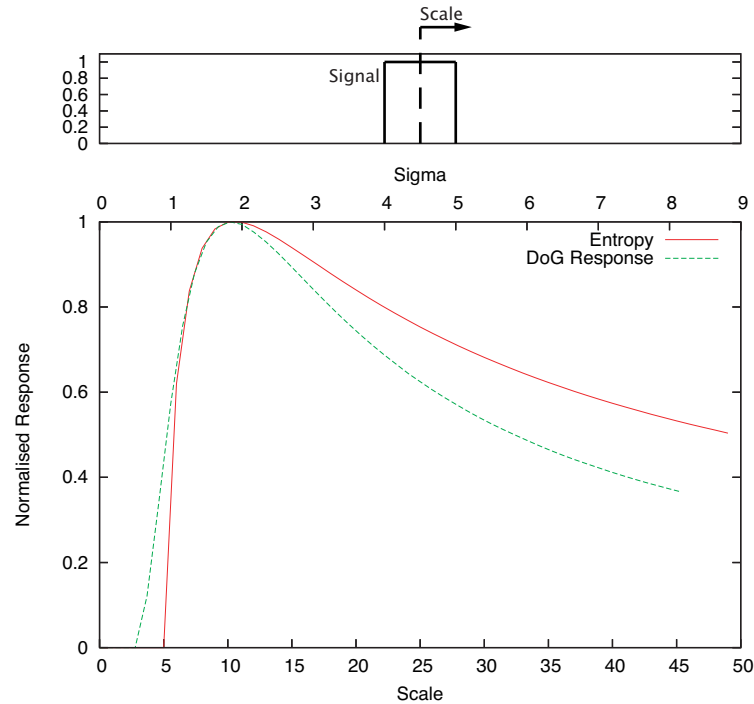


FIGURE 3.1: Entropy and difference-of-Gaussian (ratio of σ 's = 1 : 1.6, smaller σ is shown on the top x-axis) response versus scale to a one-dimensional signal as illustrated in the top diagram. The centre of the DoG and Entropy mask are kept at a constant position relative to the signal (shown by the dashed line). The graph illustrates how the response functions behave in a similar manner across scale-space

response of a difference-of-Gaussian filter is large, we would also expect the entropy taken over the same area as the filter to be large. This is illustrated in Figure 3.1. Note that the converse is not always true though: high entropy does not necessarily mean that there would be a large difference of Gaussian response.

One problem with entropy as a measure of saliency is that it is very sensitive to noise. This is especially so at small scales, where there are relatively few pixels to sample and from which to estimate the probability density function, in order to estimate the entropy. The difference-of-Gaussian is much less sensitive to noise due to the smoothing effect of the Gaussians. This is illustrated in Figure 3.2.

The remainder of this subsection is devoted to objectively comparing the stability of the two salient region detectors, and also comparing the performance of these two detectors against the baseline performance of the Harris and Stephens corner detector (Harris and Stephens, 1988).

3.2.1.1 Repeatability

We take the measure of repeatability of interest points from Schmid et al. (2000). The concept of repeatability is described below together with some results.

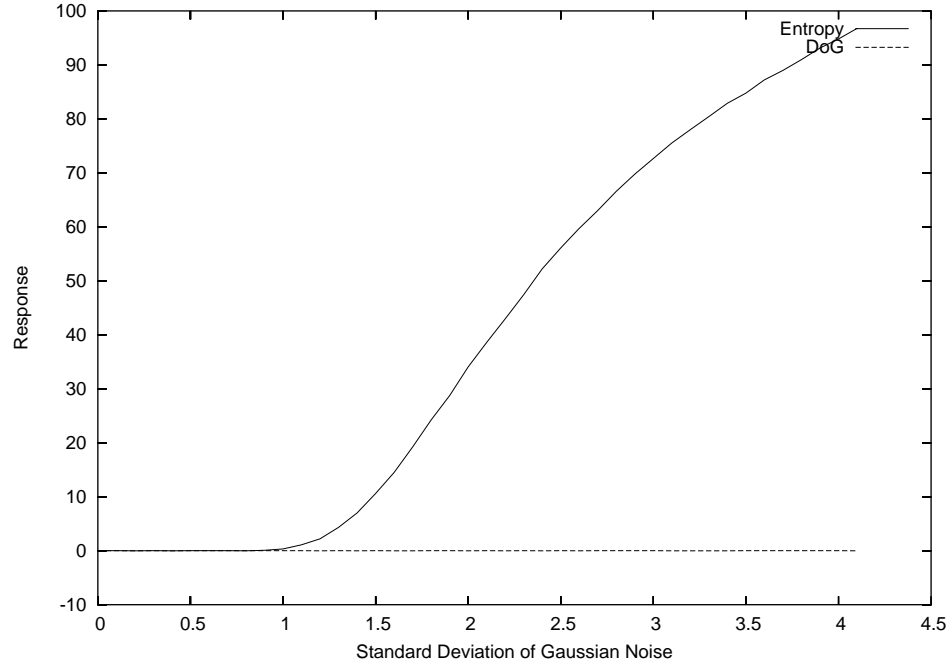


FIGURE 3.2: Response of Entropy and difference-of-Gaussian functions to a constant signal with increasing amounts of zero-mean additive Gaussian noise. The DoG response remains unshaped, whilst the Entropy response increases with noise

Repeatability Criterion. Repeatability is a measure of how independent an interest point detector is to the imaging conditions, i.e. camera parameters such as position relative to the scene, zoom, etc. 3D points detected in one image should also be detected at approximately the same corresponding locations in subsequent images. Given a point X in 3D space and two projection matrices, P_1 and P_2 , the projections of X in two images I_1 and I_2 are given by $p_1 = P_1X$ and $p_2 = P_2X$ respectively. The point p_1 , detected in image I_1 , is repeated if the corresponding point p_2 is detected in image I_2 . In order to estimate the repeatability, a unique relation between the points p_1 and p_2 has to be found. In the case of a planar scene, points in one image are related to points in a second image by a planar homography: $p_2 = Hp_1$.

The percentage of points that are repeated with respect to the total number of detected points is called the repeatability rate. In general, a point is not repeated at exactly the same position as given by Hp_1 , but in a small neighbourhood of that point. Denoting the size of the neighbourhood by ε , we can define the ε -repeatability. Interest points that cannot be observed in both images will corrupt the repeatability measure, thus only points in the common part of the scene are used to calculate the repeatability. The common part of the scene is defined by the homography, thus points \tilde{p}_1 and \tilde{p}_2 which lie in the common parts of images I_1 and I_2 are defined by $\{\tilde{p}_1\} = \{p_1 | Hp_1 \in I_2\}$ and $\{\tilde{p}_2\} = \{p_2 | H^{-1}p_2 \in I_1\}$. The set of point pairs $(\tilde{p}_1, \tilde{p}_2)$ that correspond within an ε -neighbourhood is $D(\varepsilon) = \{(\tilde{p}_2, \tilde{p}_1) | \text{dist}(\tilde{p}_2, H\tilde{p}_1) < \varepsilon\}$.

As the number of detected points in the two images may be different, the repeatability rate is defined as:

$$r(\varepsilon) = \frac{|D(\varepsilon)|}{\min(|\{\tilde{p}_1\}|, |\{\tilde{p}_2\}|)}. \quad (3.1)$$

3.2.1.2 Repeatability Results

Using the repeatability criterion, we investigated the robustness of the two salient region descriptors to image rotation and scaling. The rotation and scaling were performed digitally, using bilinear interpolation. The experiments were performed by using all of the images in the University of Washington ground-truth image data-set (University of Washington, Accessed 6/11/2003). Some example rotated and scaled test images are shown in Figure 3.3.

As a baseline with which to compare our results, we also calculated the repeatability of the well-known Harris corner detector (using a $[-2 \ -1 \ 0 \ 1 \ 2]$ kernel¹), and an improved version of the Harris detector that calculates the derivatives more precisely by replacing the $[-2 \ -1 \ 0 \ 1 \ 2]$ kernel with one calculated from the derivatives of a Gaussian ($\sigma = 1.0$).

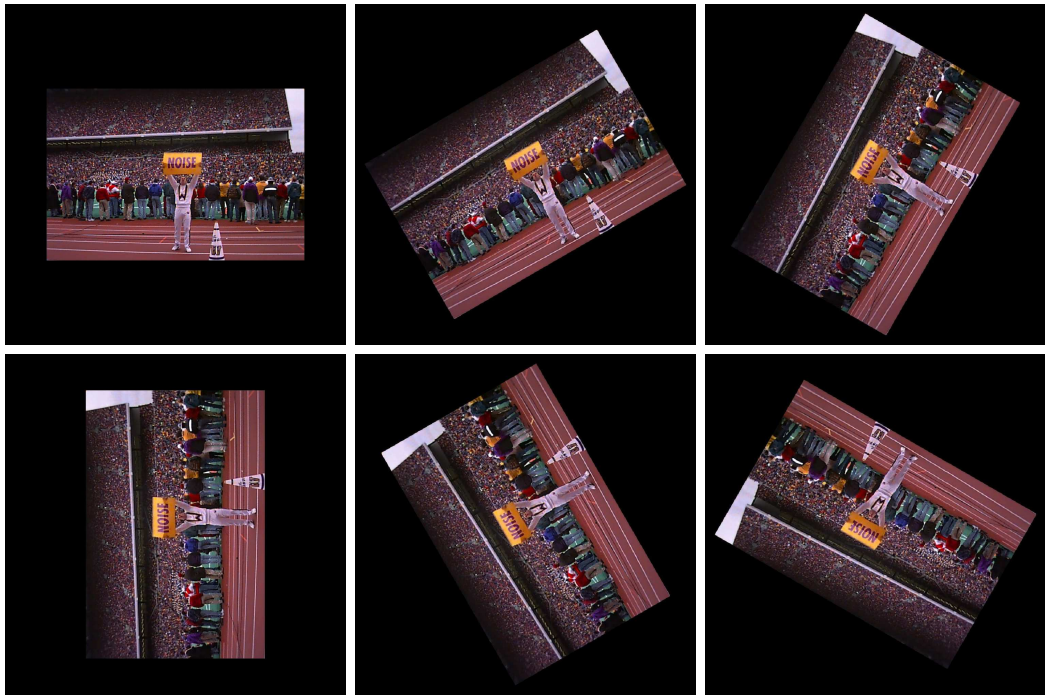
Figure 3.4(a) illustrates the results of repeatability against rotation angle, averaged over all of the images in the data-set, and Figure 3.4(b) illustrates the variation in repeatability over a range of image scales, again averaged over all the images in the dataset. The results show that the salient regions detected by finding peaks in the difference-of-Gaussian pyramid are by far the most stable to both rotation and scaling. The salient-scales algorithm performs more-or-less on a par with the Harris detector. Unfortunately, whilst the salient-scales algorithm should be robust to both scaling and rotation, in practice it is affected by discretisation of the digital raster, especially at small scales. Also, our observations have led us to believe that the clustering part of the salient scales algorithm does little to help its stability.

Recently, Kadir et al. (2004) have suggested an extension to the original Salient Scales algorithm, which include affine invariance, and also an anti-aliased sampling technique that should help with the problems of estimating the PDFs and entropy. The performance of this new detector is investigated in the next subsection.

3.2.2 DoG-Peaks and the State-of-the-Art Affine-Invariant Detectors

Recently, Mikolajczyk et al. (2005) published a comparison of six state-of-the-art affine-invariant region detectors. In addition, they fully detailed their methodology and provided the set of images which were used in their experiments. In this subsection, the

¹This kernel corresponds to a finite difference gradient estimation, which is not very robust to noise in the image.



(a)



(b)

FIGURE 3.3: A sample image from the Washington data-set showing varying amounts of in-plane rotation (a) and scaling (b) as used for the repeatability experiments.

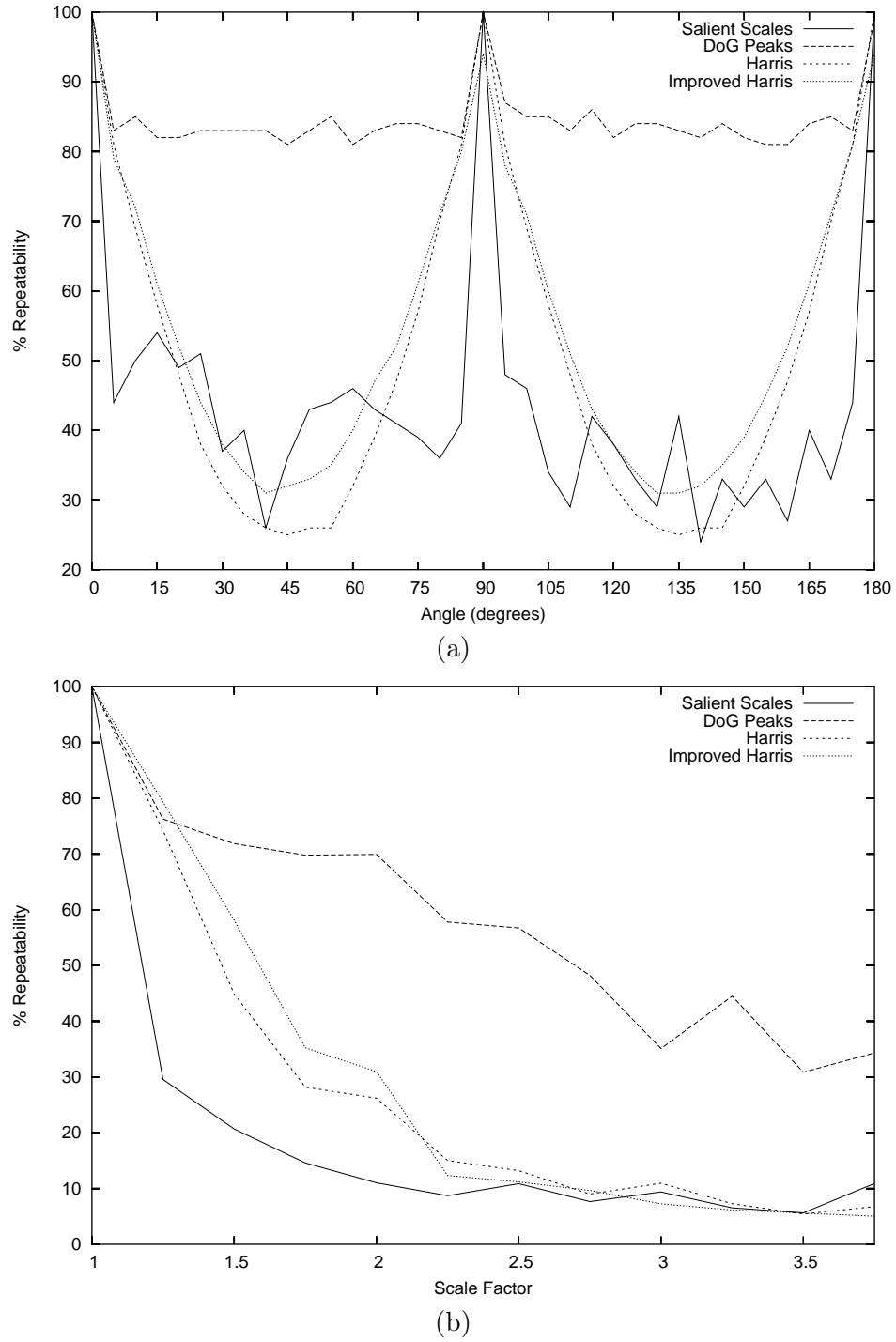


FIGURE 3.4: Repeatability rate for image rotation (a), and for scale change (b). $\varepsilon = 1.5$ in both cases

methodology and data-set are presented together with the results of running the experiments using Lowe's difference-of-Gaussian peaks detector. The performance of the difference-of-Gaussian detector is discussed with respect to the six affine-invariant detectors.

3.2.2.1 Image Data-set

The data-set proposed by Mikolajczyk et al. consists of eight sequences of six images with gradually varying photometric or geometric transformations. The data-sets are illustrated in Figures 3.5-3.9. Five different changes in imaging conditions are evaluated: viewpoint change (Figure 3.5), scale change (Figure 3.6), image blur (Figure 3.7), JPEG compression (Figure 3.8) and illumination (Figure 3.9). The viewpoint change, scale change and image blur imaging conditions are characterised by two different scene types. The first (a) is a *structured* scene with homogeneous regions with distinctive edge boundaries, whilst the second is a *textured* scene containing repeated textures of different forms.

All of the images have a resolution of approximately 800×640 pixels. In the viewpoint change test sequence the camera moves from a parallel frontal view to a view with the camera rotated by about 60° out of plane. The scale change and blur sequences were generated by modifications to the camera zoom and focus respectively. In the case of scale change, the zoom was adjusted to give a scale change of a factor of about four over the sequence. The illumination sequence was created by adjusting the camera aperture. The JPEG compression sequence was created by compressing the initial image by increasing amounts, using an image quality parameter in the encoding software. Decreasing the quality setting corresponds to more coarse-grained quantisation of the DCT coefficients in the JPEG compression algorithm.

Because each the scenes was either planar, or the camera was fixed during capture, the images in each sequence are related by planar homographies. Accurate homographies between the first image in each set and every other image in the set are provided with the data-set. The homographies have a root-mean-square error of less than 1 pixel per image pair.

3.2.2.2 Region Overlap and Repeatability

The regions detected by the affine detectors are elliptical in shape. Mikolajczyk et al. suggest the use of the *overlap error* to detect whether two regions between the original image and the transformed image correspond. Because the ellipses will all be of different sizes, or scales, the larger regions would automatically have a better chance of yielding good overlap scores. In order to make the overlap error insensitive to such scaling for the comparison of multiple region detectors, the reference region is re-scaled by a factor s to a known size (30-pixels radius in our experiments), and the target region is scaled by s .

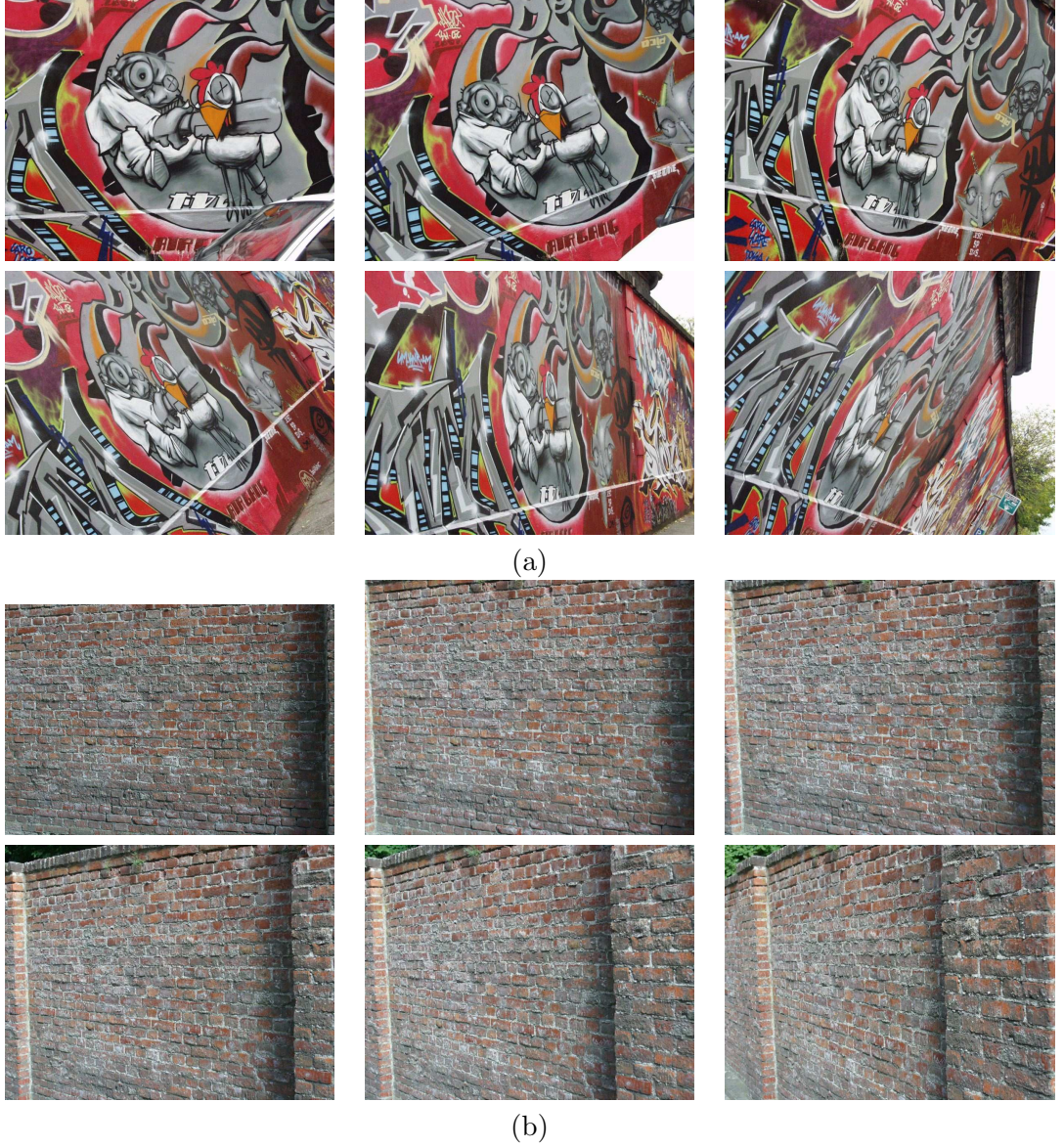


FIGURE 3.5: **Affine data-set:** Viewpoint change. (a) Graffiti sequence. (b) Wall sequence.

Mathematically, two regions are said to correspond if the overlap error, ε_O , defined as the error in the area covered by the two regions, is sufficiently small:

$$1 - \frac{R_{\mu_a} \cap R_{(H^T \mu_b H)}}{R_{\mu_a} \cup R_{(H^T \mu_b H)}} < \varepsilon_O, \quad (3.2)$$

where R_{μ} is the elliptical region defined by $x^T \mu x = 1$. H represents the planar homography between the two images. The intersection and union of the region is represented by $R_{\mu_a} \cap R_{(H^T \mu_b H)}$ and $R_{\mu_a} \cup R_{(H^T \mu_b H)}$ respectively.

Repeatability measure. The repeatability of the detector is calculated as before, using Equation 3.1, although using region matches (R_{μ_a}, R_{μ_b}) , instead of point matches

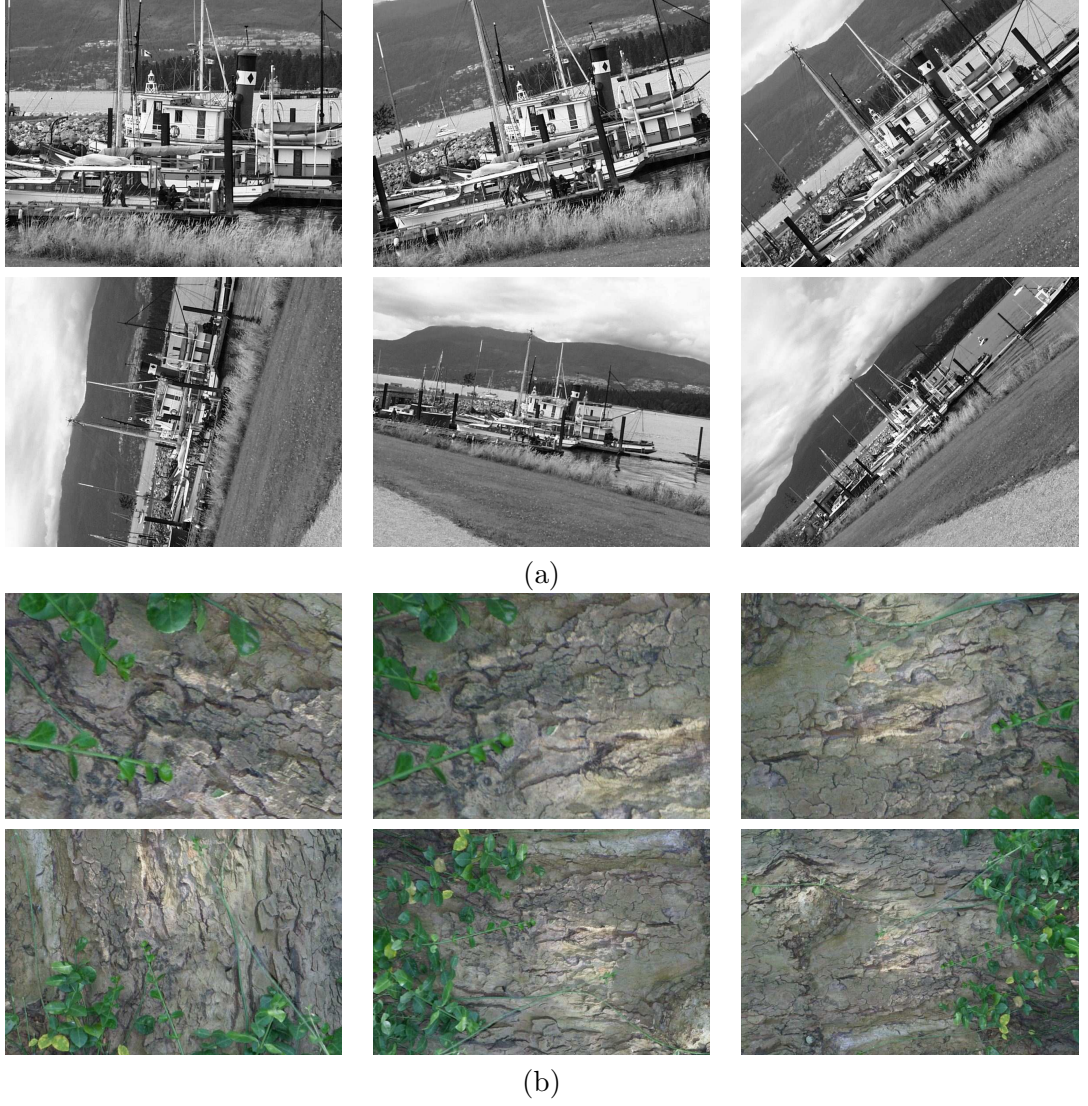


FIGURE 3.6: **Affine data-set:** Zoom and rotation. (a) Boat sequence. (b) Bark sequence.

$(\tilde{p}_1, \tilde{p}_2)$. $D(\varepsilon)$ is related to the overlap error and is defined as

$$D(\varepsilon_O) = \left\{ (R_{\mu_a}, R_{\mu_b}) \left| 1 - \frac{R_{\mu_a} \cap R_{(H^T \mu_b H)}}{R_{\mu_a} \cup R_{(H^T \mu_b H)}} < \varepsilon_O \right. \right\}. \quad (3.3)$$

In order to assess the performance of the difference-of-Gaussian detector using this framework, the circular regions found by the detector are parameterised as ellipses with equal major and minor axes. The radius of the circle/ellipse is set to the size of the standard deviation σ of the smallest Gaussian used in the difference-of-Gaussians.



(a)



(b)

FIGURE 3.7: **Affine data-set:** Image blur. (a) Bikes sequence. (b) Trees sequence.FIGURE 3.8: **Affine data-set:** JPEG Compression. UBC sequence.

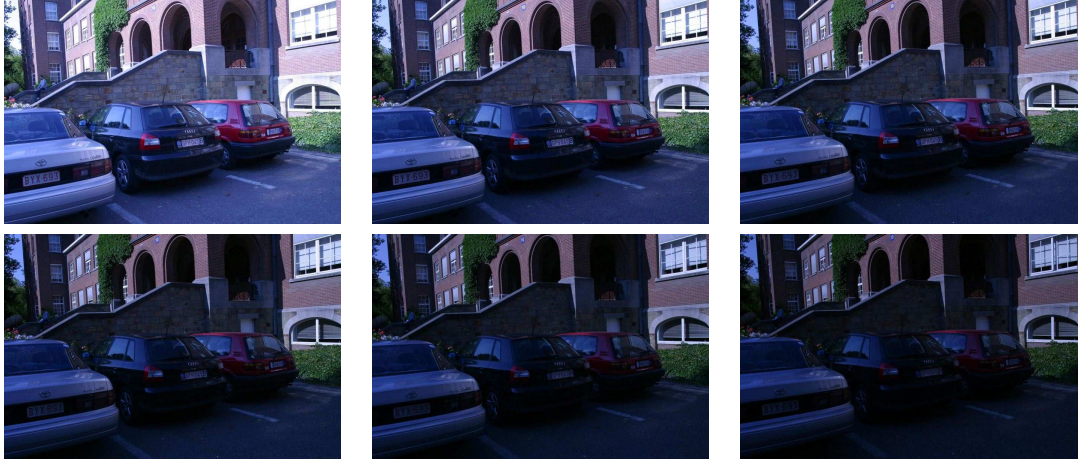


FIGURE 3.9: **Affine data-set:** Lighting change. Leuven sequence.

3.2.2.3 Matching

The evaluation of the detectors using the repeatability criterion is somewhat theoretical. It is useful to examine the detector performance from a more practical point of view, investigating the performance of the detectors in a matching scenario. By looking at the number of correct matches and the ratio of correct matches to incorrect matches, it is possible to assess performance of the detectors.

Mikolajczyk et al. (2005) suggest the use of Lowe's SIFT descriptor to describe each region. The elliptical regions are scaled up by a factor of three, and contents of each region are mapped to circular regions of 30×30 pixels in order to calculate the descriptor. Descriptors are compared using Euclidean distance.

Matching score. The *matching score* is computed as the ratio of correct matches to the smaller number of detected regions in the image pair. A match is the nearest neighbour in descriptor space. Matches are deemed as correct if the overlap error is less than 40%, or $\varepsilon_O < 0.4$. Only a single match is allowed for each region.

The matching score can be used to give an indicative idea of the distinctiveness of the features. If the matching score results do not follow the trends of the repeatability tests for a particular feature type, that means that the distinctiveness of these features differs from the distinctiveness of the other detectors.

3.2.2.4 Discussion of results

The experimental set-up described above has been applied to the difference-of-Gaussian detector in order to assess its performance with respect to the Hessian-affine, Harris-affine, MSER and Salient-affine detectors. Due to the large number of regions detected

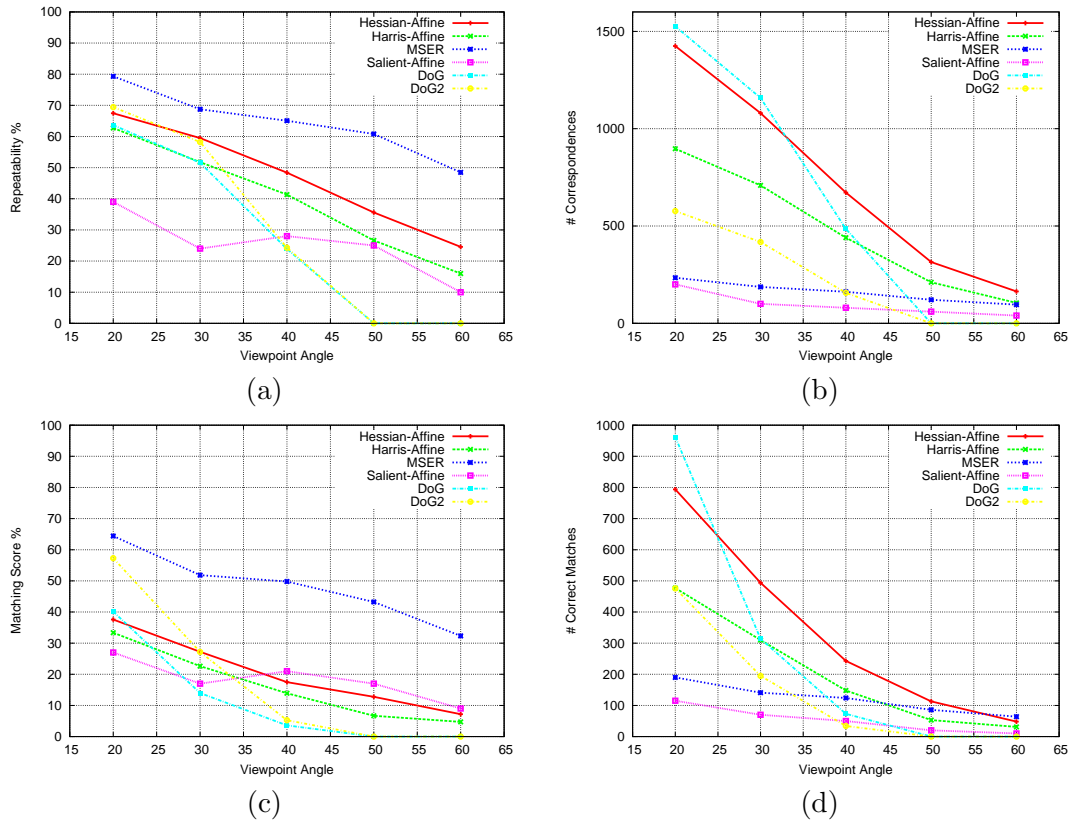


FIGURE 3.10: Viewpoint changes for the *structured Graffiti* sequence (Figure 3.5(a)). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

by the difference-of-Gaussian detector on some of the scenes, not all results were able to be generated. However, a modified version of the difference-of-Gaussian detector (referred to as *DoG2* in the graphs) was developed that rejected really small regions with sizes of less than 2 pixels. Generally speaking, this actually improved the performance of the detector because the small regions were the most likely to be affected by the transforms, and thus less repeatable and less suitable for generating descriptors which to match.

The discussion proceeds by first making some general observations about the detectors, and then looking at each of the transformations in detail. Finally some conclusions are drawn.

General observations. The first observation we make is about the computational efficiency of the detectors. In our experiments, the Harris-affine, Hessian-affine and difference-of-Gaussian detectors all took about the same time to run. The MSER detector was much faster, and the Saliency-affine detector was many orders of magnitudes slower. For example, on the first image of the Graffiti sequence (Figure 3.5(a)), the MSER detector takes under a second on average, the Hessian-affine and Harris-affine

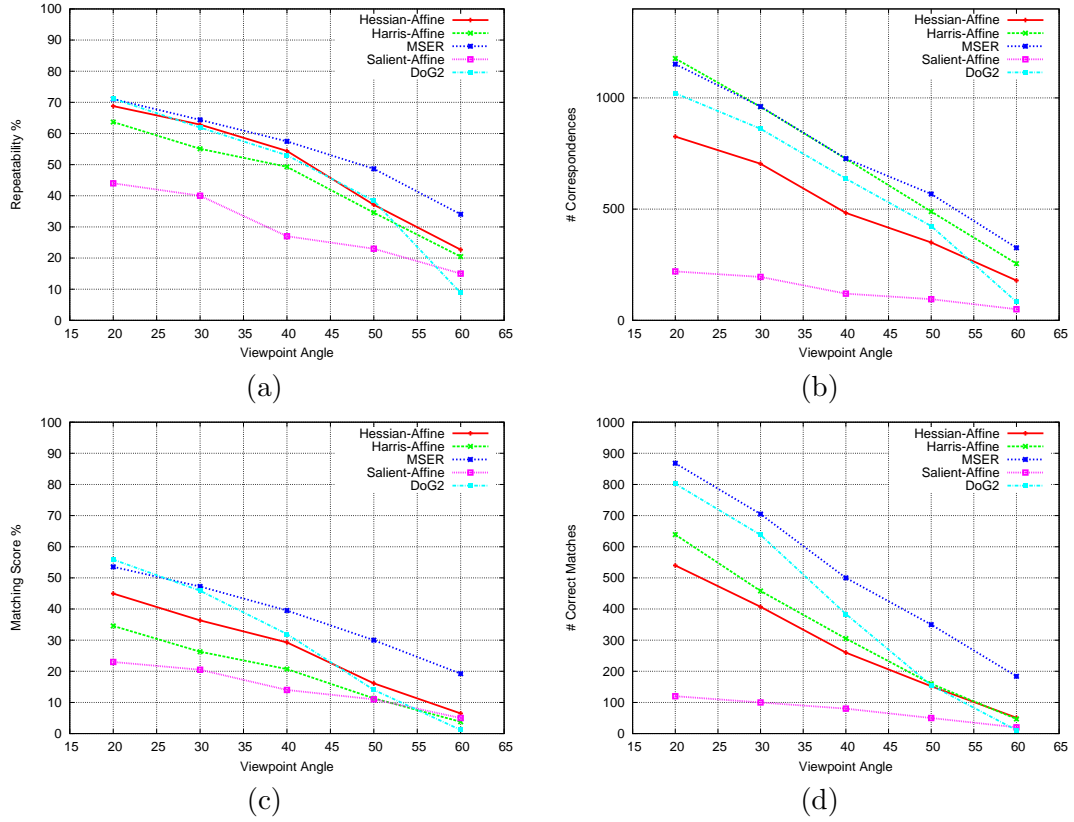


FIGURE 3.11: Viewpoint changes for the *textured Wall* sequence (Figure 3.5(b)). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

detectors both take a fraction over five seconds, the DoG detector takes about six seconds, and the Saliency-affine detector takes over an hour. It should however be noted that the codes implementing the detectors are not particularly optimised.

Our second observation is about the number of regions detected by each of the detectors. Table 3.1 shows the number of regions detected by the different detectors in the first image of the Graffiti sequence. All of the feature detectors are sensitive to the type of scene, for example, the Harris-affine detector finds almost twice as many regions in the Boat scene as in the Graffiti scene. This variability is because the detectors all respond to different features within the image. On the whole, the difference-of-Gaussian (DoG) detector finds the most features in all scene types.

Analysis of each transform. The results of the repeatability tests are shown in Figures 3.10-3.17(a) & (b). The results of the matching tests are shown in Figures 3.10-3.17(c) & (d). Ideally, the repeatability and matching score plots would have a horizontal line at 100%. As can be seen from the graphs, none of the detectors actually reaches 100% performance, and the general trend is for performance to degrade as the transform becomes more severe.

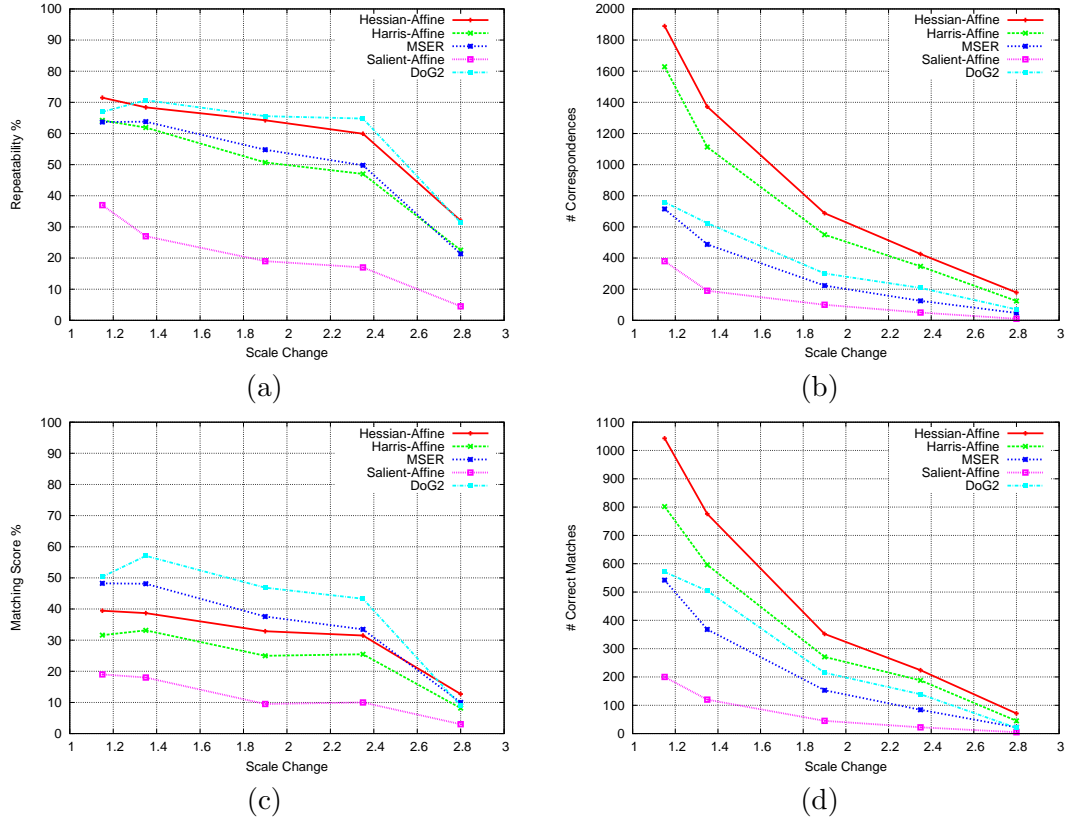


FIGURE 3.12: Scale changes for the *structured* Boat sequence (Figure 3.6(a)). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

Detector	Number of regions
Harris-affine	1758
Hessian-affine	2454
MSER	533
Salient-affine	513
Difference-of-Gaussian (DoG)	3079
Filtered Difference-of-Gaussian (DoG2)	1048

TABLE 3.1: Number of regions detected by each detector for top-left image in Figure 3.5(a)

Viewpoint change: The experimental results from the change of viewpoint in the Graffiti sequence (Figure 3.5(a)) are shown in Figure 3.10. The results from the Wall sequence (Figure 3.5(b)) is shown in Figure 3.11. The graphs show that the MSER detector performs the best on these two test sequences, both in terms of repeatability and matching score. The MSER detector performs especially well on the Graffiti sequence, where there are large amounts of homogeneous regions with very distinctive boundaries. The DoG2 detector performs fairly well for small changes in viewpoint, although it begins to fail sharply after viewpoint changes of more than $30^\circ - 40^\circ$.

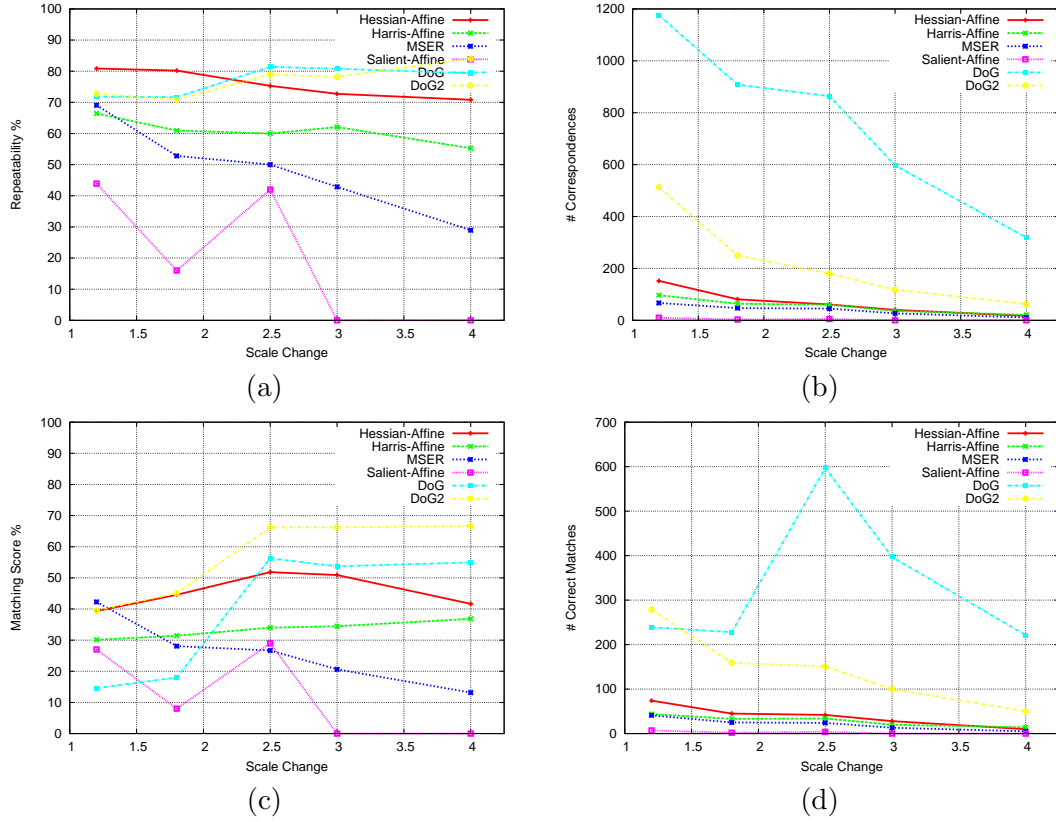


FIGURE 3.13: Scale changes for the *textured* Bark sequence (Figure 3.6(b)). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

Scale change: Figure 3.12 shows the results for the *structured* Boat sequence depicted in Figure 3.6(a) and Figure 3.13 shows the results for the *textured* Bark sequence shown in Figure 3.6(b). In both cases, the DoG2 detector is generally the best, especially at larger scale changes. The repeatability of the DoG2 detector is similar to that of the Harris-affine detector, although the matching score is as much as 10% better. These results confirm the high performance of the automatic scale selection in the Gaussian scale-space used by the difference-of-Gaussian detectors. The Salient-affine detector performs really poorly on the *textured* scene as can be seen from the very unstable repeatability and matching score results. This is due to the very low number of regions detected on scenes of this type.

Interestingly, in the *textured* sequence, the DoG and DoG2 detectors have very similar repeatability, although the number of corresponding regions for the DoG detector is much larger. However, looking at the matching score, the DoG2 detector performs much better than the DoG detector, giving credence to the earlier assertion that the smaller regions detected by the DoG detector were the least stable. It also gives some indication that the features describing the small regions are much less distinctive (which

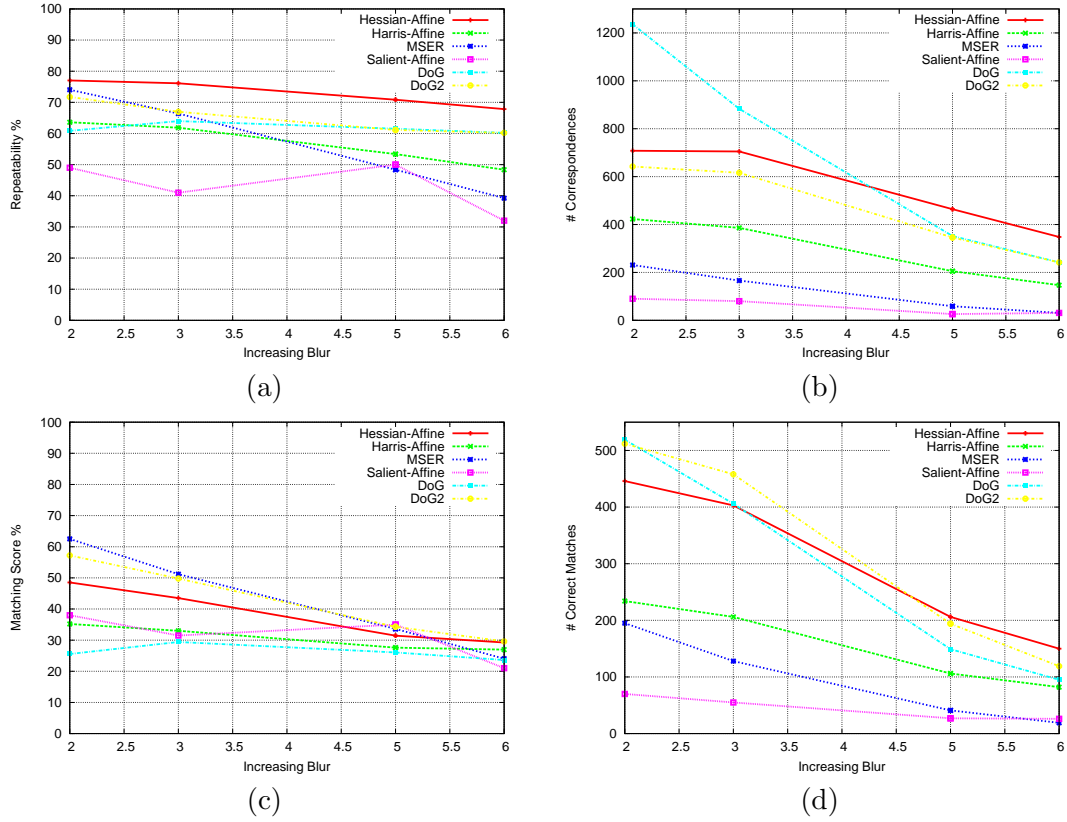


FIGURE 3.14: Blur for the *structured* Bikes sequence (Figure 3.7(a)). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

is obvious, because of the smaller number of pixels used to sample in the creation of the descriptor).

Blur: The results of the blur experiments from the Bikes (Figure 3.7(a)) and Trees (Figure 3.7(b)) sequences are shown in Figures 3.14 and Figure 3.15 respectively. On the whole, the results for the blur experiment are better than those for the scale- and viewpoint-change experiments. With the exception of the MSER detector, all the detectors have almost horizontal repeatability and matching score curves, indicating that the detectors are less sensitive to increasing blur than the other factors. The MSER detector is more sensitive to blurring because as blur is increased, the region boundaries become smoother and the segmentation becomes less accurate.

In the *structured* Bikes scene, the Hessian-affine and DoG2 detectors have the best performance; The Hessian-affine detector is more-or-less consistently 10% better than the DoG2 detector in terms of repeatability, but the DoG2 detector is better in terms of matching score performance. In the *textured* scene the DoG2 and Hessian-affine detectors perform best and have very similar repeatability performance, whilst all of the detectors have a very similar matching score performance within a 10% band. This

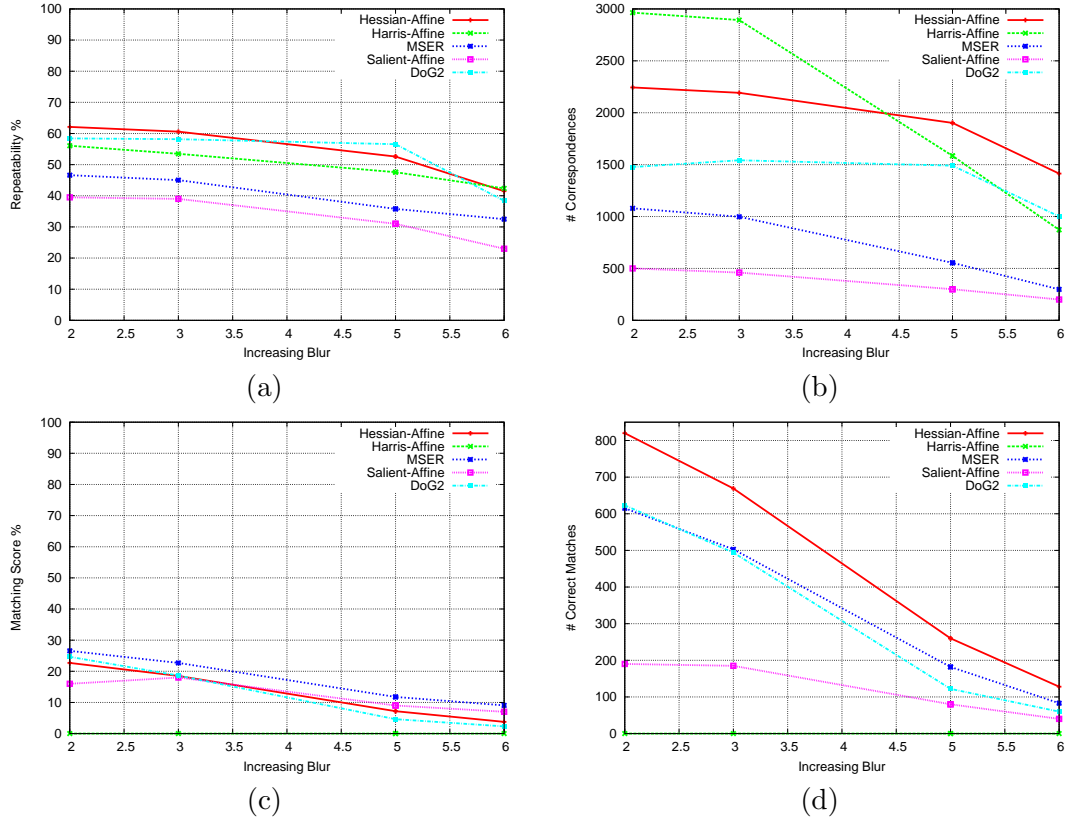


FIGURE 3.15: Blur for the *textured Trees* sequence (Figure 3.7(b)). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

implies that the features of each detected region are quite similar, hence a large number of mismatches, and low matching score. Looking at the scene, it can be seen that there are a large number of local structures that are barely distinguishable.

JPEG compression: Figure 3.16 shows the results from the JPEG compression experiment with the UBC sequence shown in Figure 3.8. The JPEG compression experiment is interesting because as the compression ratio is increased, more and more information is lost, but also, new artefacts are introduced. The results show that the Hessian-affine and Harris-affine detectors clearly have the best performance for this type of scene, followed by the DoG2 detector. All of the detectors show the same trends with increasing compression in terms of both repeatability and matching score.

Illumination change: Figure 3.17 shows the effect of lighting change as illustrated in the Leuven sequence shown in Figure 3.9. All of the detectors show very good robustness to lighting change as the curves are almost horizontal. Overall, the MSER detector performs best for the scene, followed by the DoG2 detector.

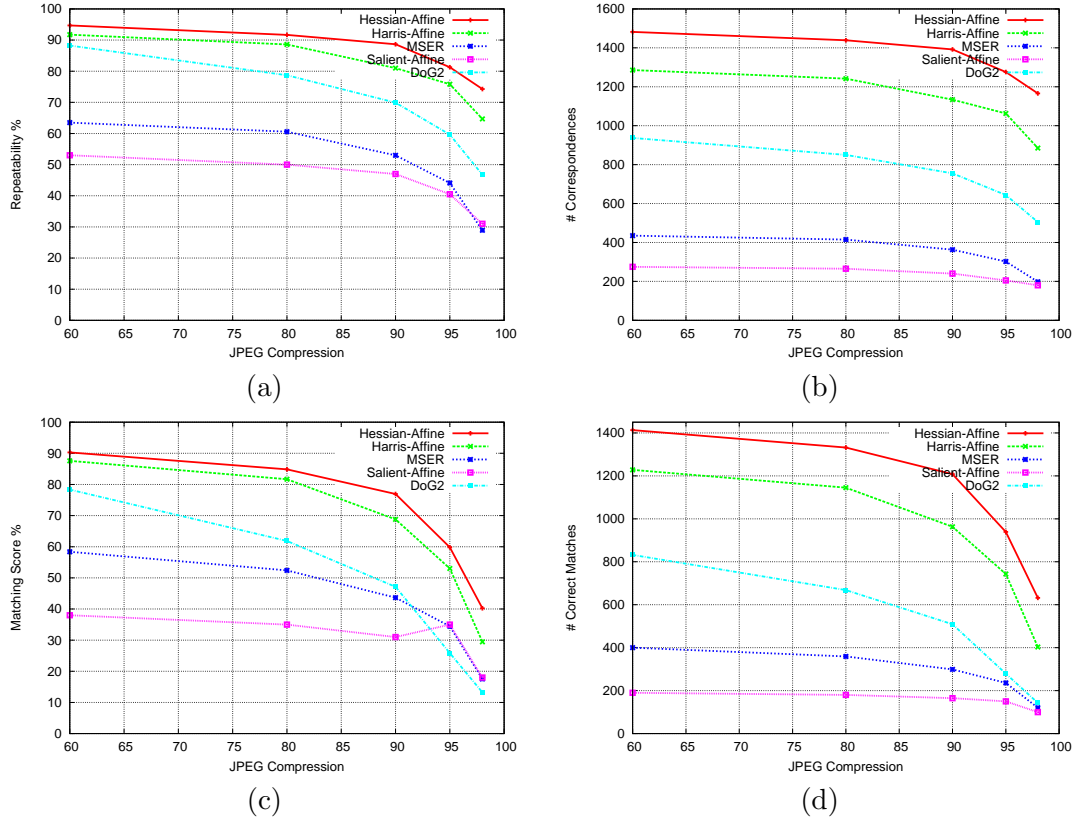


FIGURE 3.16: Increasing JPEG compression for the UBC sequence (Figure 3.8). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

Conclusions. The Saliency-affine detector was the worst performing in almost all of the tests. The MSER, Hessian-affine and DoG2 detectors obtained the best repeatability and matching scores for most experiments. On the whole, with the exception of viewpoint changes of more than $30^\circ - 40^\circ$ the DoG2 performed the most consistently across the different scene types. Generally the matching plots looked similar to the repeatability plots, albeit with lower values. As previously mentioned, this indicates that the regions have sufficient distinctiveness to be matched automatically. When the relative order of the detectors changes between two plots, it implies that the regions found by some of the detectors are not distinctive enough, and a large number of mismatches occur.

The viewpoint change was the most difficult transformation for the detectors to cope with, followed by the scale change. The repeatability results for most of the detectors were generally consistent across each of the sequences. However, in the blur sequence of Figure 3.14, the MSER detector repeatability performance degrades much more rapidly than with the other detectors. The difference-of-Gaussian (DoG and DoG2), Hessian-affine and Harris-affine detectors all provide several times more corresponding regions than the other detectors.

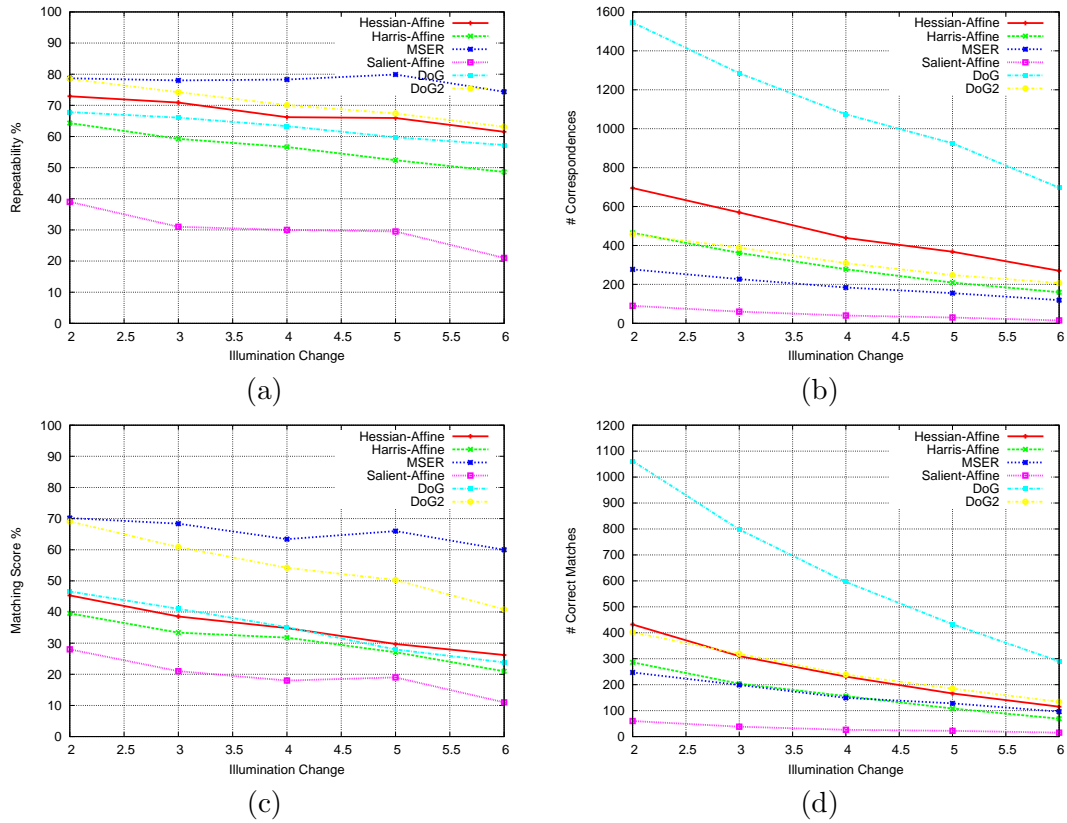


FIGURE 3.17: Decreasing Illumination for the Leuven sequence (Figure 3.9). (a) Repeatability score (regions normalised to a radius of 30 pixels, 40% overlap error). (b) Number of corresponding regions. (c) Matching Score with the SIFT Feature. (d) Number of correct nearest-neighbour matches using the SIFT feature.

3.3 A Simple Local Colour Descriptor

The SIFT descriptor has been shown to be a very robust descriptor of local image structure (Mikolajczyk and Schmid, 2003), however it only calculates structure from the intensity channel of the image. Previously, it has been shown that content-based retrieval can be improved by using some form of colour descriptor. The obvious approach to calculating a simple colour descriptor is to take a histogram of the pixel values over the colour space (Swain and Ballard, 1991). However, this approach has problems when used in combination with salient regions; because salient regions are relatively small, they have few pixels from which to sample to generate an accurate histogram.

A different approach is suggested here, inspired by the Multimodal Neighbourhood Signature (MNS) algorithm developed by Matas et al. (2000). The MNS algorithm uses the mean-shift algorithm to cluster pixels in colour-space, in order to determine the dominant colours within a local neighbourhood. The advantage using the mean-shift algorithm is that it doesn't require prior knowledge of the number of clusters. By applying the mean-shift algorithm to pixels within each salient region, it is possible to estimate

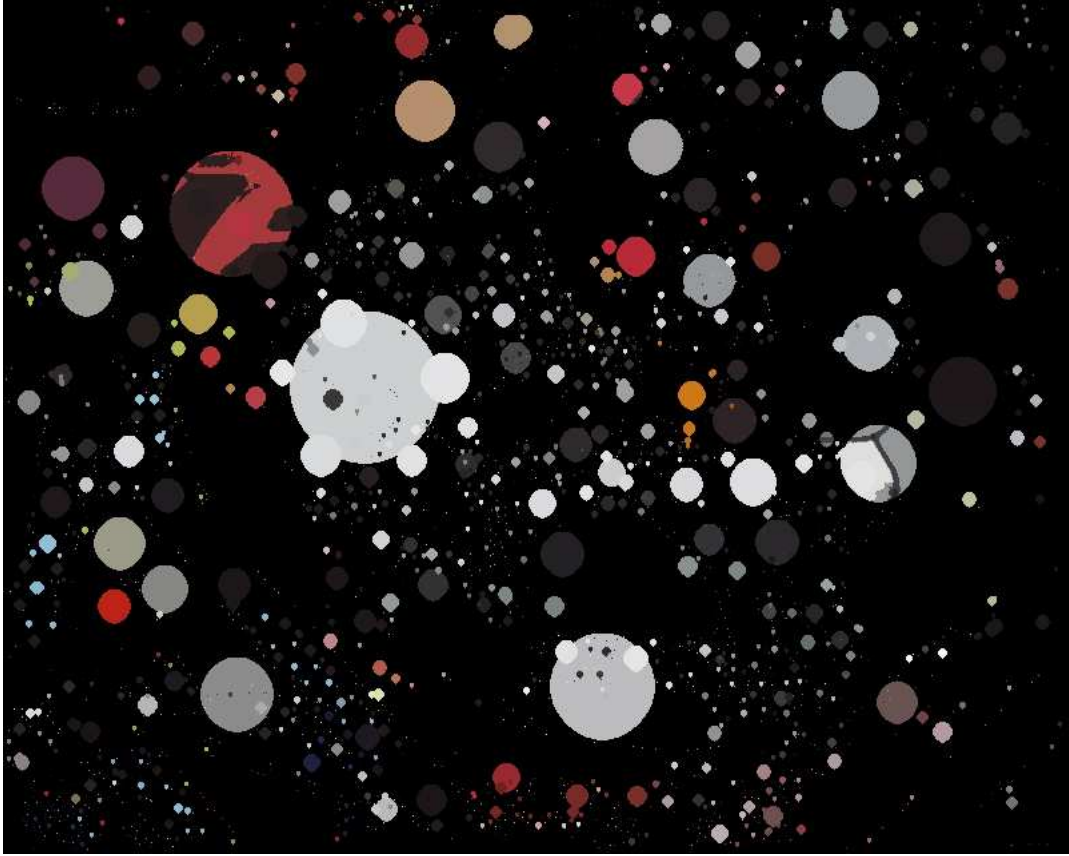


FIGURE 3.18: The dominant colour descriptor applied to DoG regions on the first image of the Graffiti sequence. Each region is shaded by mapping the original pixels to their dominant colours. The illustrated regions radii is the variance of the smaller of the Gaussians in the DoG.

the modes of the colour-space within the region, and thus we have a way of indexing regions based on their colour-modality.

In our implementation of the algorithm, colours are clustered in RGB space, however it is possible to transform the dominant colours into a different colour-space for indexing purposes. For example, HSI or intensity-normalised RGB space may be more practical for retrieval. This issue is discussed in more depth in the following chapters. In practice when the algorithm is applied to salient regions, the majority of regions are represented by a unimodal colour distribution. Whilst this limits the distinctiveness of the regions, this is not a problem as the colour descriptor is most often used together with another descriptor, such as the SIFT descriptor. Figure 3.18 illustrates the effect of applying the colour descriptor to the first image in the Graffiti sequence in Figure 3.5(a). The figure shows each of the salient regions detected by the DoG detector filled with their dominant colours.

3.4 Summary

This chapter has discussed a number of factors affecting the use of saliency in robust retrieval scenarios. The chapter has presented two in-depth comparisons of different saliency detectors, in particular looking at the performance of the difference-of-Gaussian detector that will be used in the remainder of this thesis. The difference-of-Gaussian detector has been shown to be very robust to a number of different geometric- and photometric-transforms that may reasonably be expected to occur in a retrieval scenario. The detector does have some performance problems when the scene is subjected to extreme out-of-plane viewpoint rotation. However, as explained in the later chapters, this is not something we expect to have to deal with in our retrieval scenarios.

The final part of the chapter describes a simple local colour descriptor that will be used in later chapters to augment retrieval using the SIFT descriptor.

Chapter 4

Image Retrieval using Salient Region Descriptors

“Often the search proves more profitable than the goal.”

E. L. KONIGSBURG

“Getting information off the Internet is like taking a drink from a fire hydrant.”

MITCHELL KAPOR

This chapter investigates how image retrieval can be performed using local descriptors of salient regions. The first section of the chapter details an investigation, first presented by Hare and Lewis (2004), into the use of the difference-of-Gaussian detector for retrieval and verifies previous results that showed that retrieval using salient regions could outperform global descriptors. The second part of the chapter looks at how the vector-space and Latent Semantic Indexing text retrieval techniques can be applied to images using salient region descriptors.

4.1 Basic Model

In previous work by Sebe et al. (2003), the use of salient point detectors for content-based retrieval was shown to have better performance than when using global descriptors. In this section we describe a new metric for measuring the performance of content-based retrieval based on salient regions, and illustrate it with some preliminary results that show that the performance when using salient regions is indeed better than when using global descriptors. Previous work by Hare and Lewis (2003) showed that salient regions

detected by the original Scale-Saliency algorithm could be used for retrieval and sub-image matching using simple grey-level histogram descriptors.

In order to facilitate the testing of the the use of salient regions for content-based retrieval, we have developed a system that returns the N closest matches to a given query image. The system enables queries to be made using either global descriptors or a descriptor based on salient regions. Following Sebe et al., we fix the number of salient regions to 50 per image. In the case of global descriptors the distance D_E between two images I_1 and I_2 is given by the Euclidean distance

$$D_E(\mathbf{F}_1, \mathbf{F}_2) = |\mathbf{F}_1 - \mathbf{F}_2| = \sqrt{\sum_{i=1}^K |\mathbf{F}_{1_i} - \mathbf{F}_{2_i}|^2} \quad (4.1)$$

between the feature descriptors \mathbf{F}_1 , and \mathbf{F}_2 , where K is the number of elements in the feature descriptors. In the case of matching using salient regions, the distance between two images is given by a linear summation of the closest matching feature vector in the second image for each feature vector in the first image. Denoting the set of M feature vectors in images I_1 and I_2 as $\{\mathbf{F}_1\}$ and $\{\mathbf{F}_2\}$ we define

$$D_{\text{salient}}(\{\mathbf{F}_1\}, \{\mathbf{F}_2\}) = \sum_j^M \min_k (D_E(\{\mathbf{F}_1\}_j, \{\mathbf{F}_2\}_k)), \quad (4.2)$$

where $\{\mathbf{F}_1\}_j$ refers to the j th feature vector of image I_1 and $\{\mathbf{F}_2\}_k$ refers to the k th feature vector of image I_2 .

4.1.1 Semantic Relevance

The problem with global descriptors is that they cannot fully describe all parts of an image having different characteristics. The use of salient regions aims to avoid this problem by developing descriptors that do capture the characteristics of each part of the image. Given this aim, it should not be unreasonable to expect that an image description generated from salient regions will be *better* than an image described wholly by a global descriptor. In order to test this we have developed a metric that uses semantically marked images as ground-truth against the results from our retrieval system.

The University of Washington Ground Truth Data-set (University of Washington, Accessed 6/11/2003) contains 697 images that have been semantically annotated. For example an image may have a number of labels describing the image content, such as “trees”, “bushes”, “clear sky”, etc. Figure 4.1 shows some sample images and annotations from the data-set. Given a query image with a set of labels, we should expect that the images returned by the retrieval system should have the same labels as the query



Clear Sky, Tree, Bush,
Grass



Beach, Sky, Ocean, Tree



Clear Sky, Building,
Ground, People



Clear Sky, Building, Tree,
Leafless Tree, Grass,
People, Sidewalk



Overcast Sky, Building,
Tree, Flower, Car, Pole



Tree, Tree, Leafless Tree,
Bush, Building, Street,
Sidewalk, Overcast Sky



Tree, Leafless Tree, Grass,
Overcast Sky



Tree, Building, Grass,
Sidewalk, Pole, People,
Clear Sky



Stadium, Stand, People,
Football Field, Track,
Banner, People, Band,
Line



Water, Building, Sailboat,
Sky



Clear Sky, Tree, Water



Monkeys, People

FIGURE 4.1: Sample images and their annotations from the Washington Ground Truth Image Database

Feature Type	rank-1 Result Image		Averaged Top 5 Result Images	
	DoG Peaks	Global	DoG Peaks	Global
RGB Histogram	42.1%	37.6%	51.0%	45.6%
HSI Histogram	45.2%	36.9%	50.4%	49.6%
Mono Histogram	31.6%	36.9%	42.3%	45.0%
HU Moment	41.1%	22.6%	52.4 %	39.5%
RGB Colour Moment	33.7%	24.1%	41.9%	35.4%
HSI Color Moment	34.9%	30.2%	43.5%	40.5%

TABLE 4.1: Averaged Semantic Relevance for queries based on the rank-1 result image and the closest 5 result images

image. Let A be the set of all labels from the query image, and B be the set of labels from a returned image. We then define the semantic relevance, $R_{semantic}$, of the query to be:

$$R_{semantic} = \frac{|A \cap B|}{|A|} \quad (4.3)$$

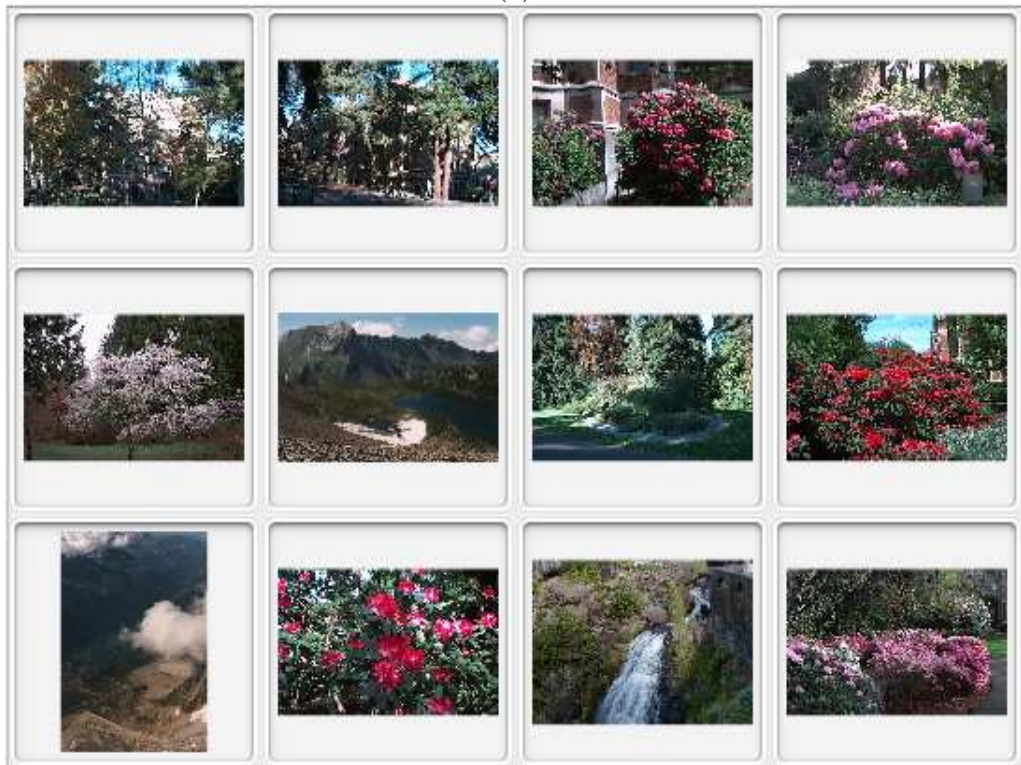
This implies that if all the labels in set A exist in set B then the semantic relevance will be 100%, and if only half of the labels in set A exist in set B then the semantic relevance will be 50%.

4.1.2 Results

We used all of the semantically marked images from the Washington data-set to form our test set. Taking each image in the test set in turn as a query, we calculated the distance to each of the other images in the test set using a range of feature types. We then calculated the semantic relevance for the rank one image (the closest image, not counting the query image), and we also calculated the averaged semantic relevance over the closest 5 images. The results of this are shown in Table 4.1. The table shows that the use of salient regions does indeed produce better semantic relevance than using global descriptors, although we believe that there is still scope for improvement of the semantic relevance from the salient regions. We believe that using a single feature type to describe a salient region (or indeed the whole image) is not sufficient. For example, the RGB histogram that represents a “blue sky” semantic label may be very similar to the histogram representing the “water” label. In our future work we hope to show it is possible to improve the semantic relevance of queries using salient regions by fusing multiple feature descriptors. Figure 4.2 illustrates the differences between a query based on a global RGB-Histogram descriptor, versus multiple RGB-Histogram descriptors based around salient regions found from the peaks in the difference-of-Gaussian pyramid.



(a)



(b)

FIGURE 4.2: Example Retrieval: (a) shows the results of a query using the Difference of Gaussian salient region method, and (b) shows the results of the same query with the Global method. In both cases, RGB Histograms are used as the feature descriptor and the first image shown is the query image

4.1.3 Discussion

This section has demonstrated that by using salient regions to generate image descriptors, it is possible to get better retrieval performance than with using global descriptors alone. However, the approach does have some limitations; firstly it is expensive - fifty times more histogram comparisons need to be performed using the salient region method, corresponding to the cost of comparing all of the individual region histograms versus a single histogram per image. Secondly the choice of fifty regions per image is somewhat arbitrary, and certainly limits the robustness of the approach. The following sections of this chapter discuss an alternative approach which solves both of these problems.

4.2 Text Retrieval Approaches

Recent work by Sivic and Zisserman (2003) on video and slightly earlier work by Westmacott and Lewis (2003), showed a new approach to object matching within images and video footage. The approach was based on an analogy with classical text retrieval using a vector-space model. This section shows how this analogy can be applied to content based retrieval from still frames using local descriptors generated from salient regions.

4.2.1 Applying Text Retrieval Techniques to Image Retrieval using Salient Regions

In this section, the ideas and methods described above for text retrieval are taken and applied to image retrieval. The analogy used is that an image is a document, and consists of multiple terms, or ‘visual’ words. In the previous chapters, the use of saliency as a means to build image descriptions was discussed. In order to build the ‘visual’ words for an image, it is suggested that each word is formed from a local description of the image in a salient region.

4.2.1.1 Building visual words: Vector Quantisation

One immediately obvious problem with taking local descriptors to represent words is that, depending on the descriptor, there is a possibility that two very similar image patches will have slightly different descriptors, and thus there is a possibility of having an absolutely massive vocabulary of words to describe the image. A standard way to get around this problem is to apply vector quantisation to the descriptors to quantise them into a known set of descriptors. This known set of descriptors then forms the vocabulary of ‘visual’ words that describes the image. The process is essentially the equivalent of stemming, where the vocabulary consists of all the possible stems.

The next problem is that of how to design a vector quantiser. Sivic and Zisserman (2003) selected a set of video frames from which to train their vector quantiser, and used the k -means clustering algorithm to find clusters of local descriptors within the training set of frames. The centroids of these clusters then became the ‘visual’ words representing the entire possible vocabulary. The vector quantiser then worked by assigning local descriptors to the closest cluster.

In the QMNS algorithm of Westmacott and Lewis (2003), RGB colour-space was quantised by splitting into regular hypercubes. The RGB-space was split into 4 bins per dimension, resulting in 64 bins total. These 64 bins correspond to a ‘visual’ vocabulary of 64 terms. Westmacott and Lewis also indexed bi-modal colour distributions as RGBRGB pairs and tri-modal distributions as RGBRGBRGB triples as 4096- and 262144-term vocabularies respectively.

The two approaches outlined above work well in different situations. The clustering-based approach is able to cope with very high-dimensional feature vectors, such as in the 128-dimensional SIFT features. The second approach only works well in low-dimensional spaces; for example, if we were to split each dimension of the SIFT features into 4 bins, we would have a vocabulary of $4^{128} \approx 10^{77}$ terms, which would be far too big to handle practically.

In this work, both approaches were used. In order to create vocabularies for the SIFT descriptors, the batch k -means clustering algorithm was used: vocabularies were created for each of the image-sets by randomly sampling 100,000 SIFT features and clustering for a number of different k values with randomly chosen start points. The clustering was performed a number of different times for each k value in order to select the best vocabulary. Each image in the image-sets then had its SIFT descriptors quantised by assigning the descriptor to the closest cluster. In order to create visual terms from the dominant colour descriptor described in 3.3, the second quantisation approach was used. Instead of indexing the raw RGB values, the colours are converted to Hue and Saturation values and these are quantised. This is to enable partial illumination invariance within the descriptors, as well as make the colour-space more like the perceptual space. The Hue and Saturation values are quantised into 60 terms by binning the Hue into 30° segments, and the Saturation into 5 bins, as shown in Figure 4.3. Colour pairs are represented as terms in a 3600-term vocabulary. In order to keep the vocabulary size down, salient regions with more than two dominant colours are represented by the two most-dominant colours within the region.

Zipf Again. Given the ubiquity of Zipf’s law in natural languages, it is interesting to see if it holds for the pseudo-artificial vocabulary of visual words created by the vector quantisation of feature vectors. It would also be interesting to investigate whether Zipf’s law can be used in choosing the optimal size of vocabulary. Figure 4.4 illustrates the

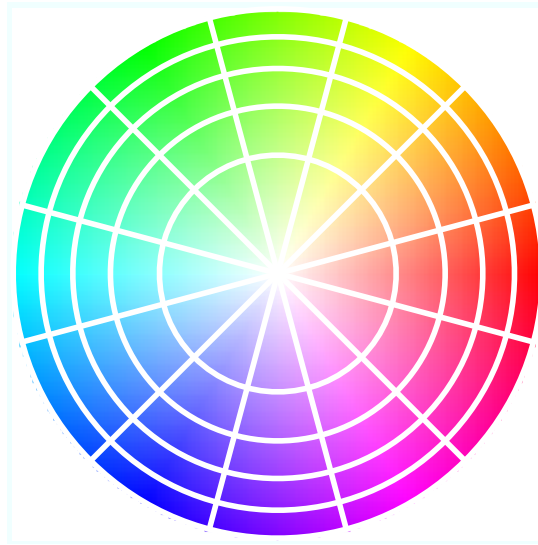


FIGURE 4.3: Illustration of how the hue and saturation are quantised to form a vocabulary of colour ‘visual’ terms. Each segment of the colour wheel represents a ‘visual’ term.

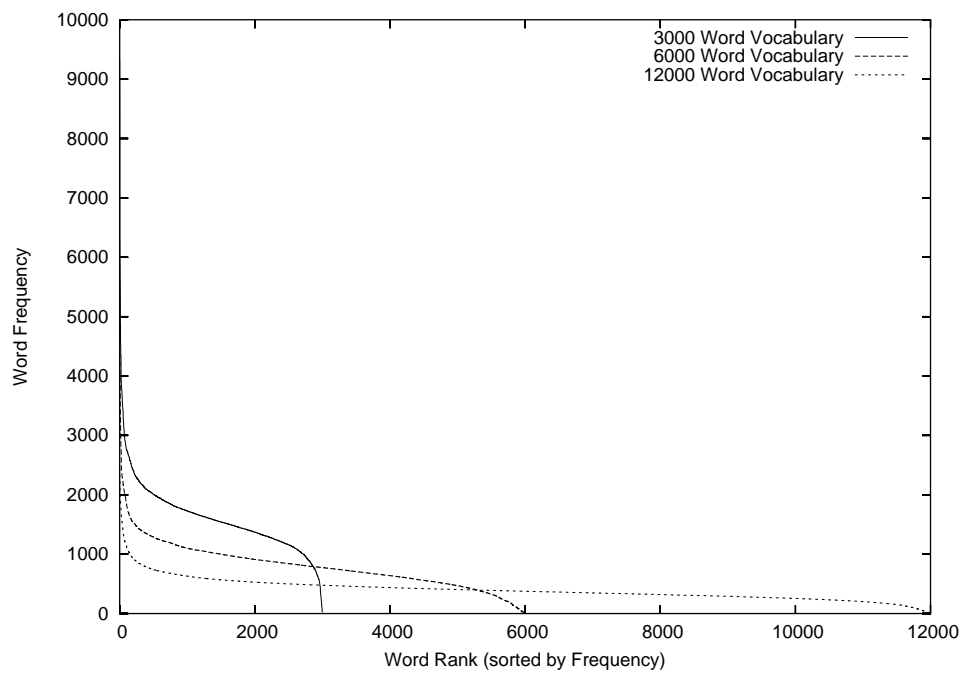


FIGURE 4.4: Rank-frequency curves for the ‘visual’ words of varying size vocabularies. The curves are generally Zipfian in nature, but the smaller vocabulary shows a large drop-off at its tail, possibly indicating that the vocabulary is too small.

rank-frequency curves calculated over the entire Washington data-set using a range of vocabulary sizes. It is interesting to note that each of the curves is approximately Zipfian, although the tail end of the smaller vocabulary curves has a noticeable non-Zipfian drop-off. A question arises as to whether this could be an indication that the vocabulary is somehow deficient, due to a lack of more descriptive visual terms.

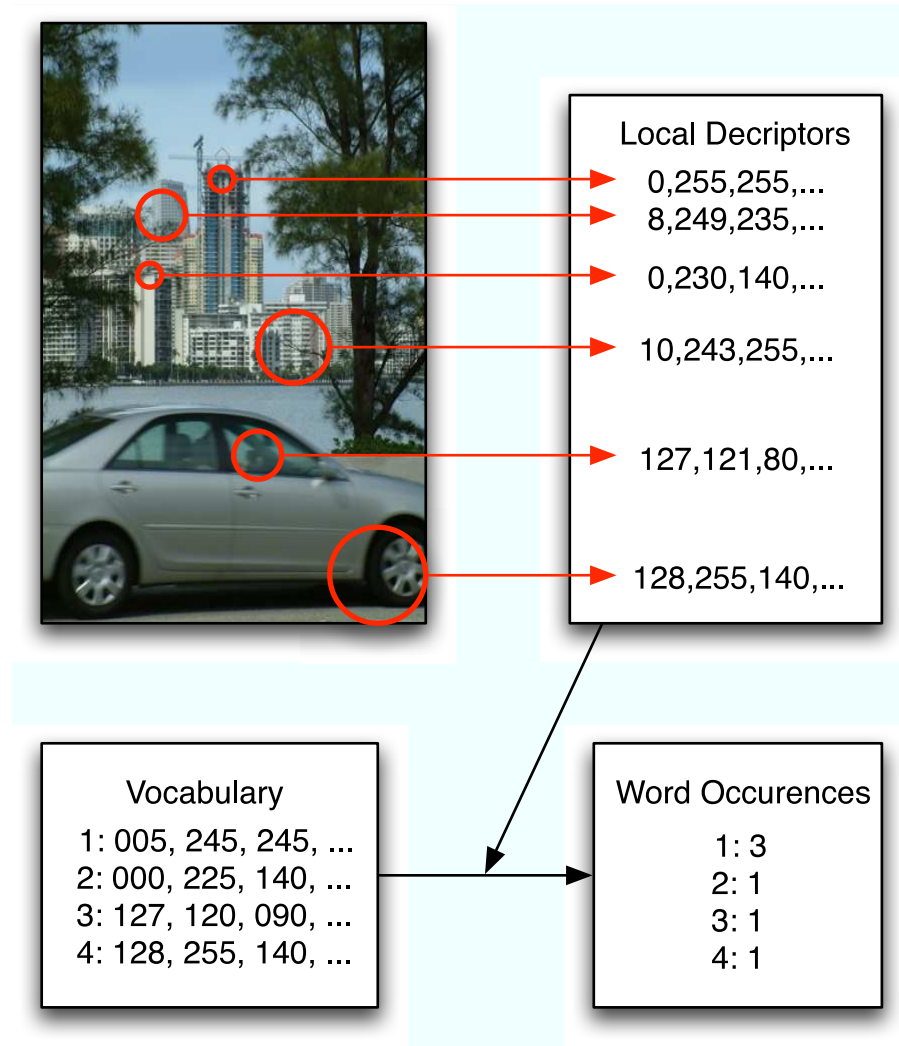


FIGURE 4.5: Generating vectors of occurrences of 'visual' terms from an image.

4.2.1.2 Image Retrieval based on visual words

Given the lists of 'visual' words for each image in the data-set, the next stage is to calculate a word-frequency vector to represent each image. The overall process of getting from an image to a vector of word occurrences is shown in Figure 4.5.

The Classical Approach The *tf-idf* weighting (Equation 2.1) is used to weight each element of the word-frequency vector. In order to perform actual retrieval, a query vector, \mathbf{V}_q is constructed from the query image, and all the documents in the database are ranked by the normalised scalar product (Equation 2.2) between the query vector \mathbf{q} and each document vector \mathbf{V}_d .

Stop Lists and Spatial Consistency. As with text retrieval, it is possible to apply a stop-list analogy to the 'visual' words. Currently, this has not been implemented,

however, it is easy to see that by adding the most common ‘visual’ words to a stop-list, much of the noise from non-discriminatory words will be removed.

It is also possible to apply constraints based on the spatial arrangement of the salient regions. This is akin to text-search methodologies where the rank of the document is increased if the query terms occur close together in that document. In terms of images, the spatial arrangement could be rigid, and tested by checking for consistent homographies between clusters of matches as in Chapter 5. Alternatively this could be measured loosely by just requiring that neighbouring matches in the query lie in a surrounding area in the retrieved image.

The Latent Semantic Indexing Approach The LSI approach to text retrieval can also be applied to image retrieval using salient regions. Given the list of ‘visual’ words for each image in the data-set, it is possible to construct the term-document matrix, apply log-entropy weighting and decompose into subspace, just as one would for text documents.

4.3 Evaluation Techniques

In order to test the performance of these two retrieval techniques, a comprehensive evaluation has been performed. Visual words from both the SIFT features and dominant colour descriptor have been tested separately, as well as by combining them into a single vector by appending the word occurrence vectors to one-another. Both the plain vector space approach and the LSI approach were tested using unweighted word occurrence vectors in addition to Log-Entropy and TF-IDF weighted vectors.

4.3.1 Data-sets

Two separate image data-sets were used for the evaluation. Firstly, the Washington Ground Truth data-set, introduced earlier, was used. This data-set consists of 697 medium-resolution images of approximately 750×500 pixels. The second data-set consists of 5000 low-resolution images from the Corel stock photo collection. Each image in the Corel collection measures about 192×128 pixels. Both of the image sets have ground-truth annotations for each image which can be used for benchmarking purposes.

The original semantic labels used for marking up the images in the Washington database are in some ways deficient because they use no predefined ontology or vocabulary; For example, some of the images have a “Garbage Can” label, whilst others have a “Trash Can” label. The measure of semantic relevance has no way of knowing these terms have the same meaning. We have applied a smaller, fixed vocabulary, to give a better indication of how semantically relevant one image is to another.

4.3.2 Precision, Recall and Semantic Relevance

In addition to comparing the image retrieval algorithms through the semantic relevance measure (c.f. Equation 4.3), we would also like to plot precision-recall curves. In order to do this, we need to know whether a particular target image is relevant to the query. Using the semantic relevance measure, above, we define the relevance of each image, $V_{n,Z} \in \{0, 1\}$, to be

$$V_{n,Z} = \begin{cases} 0 & \text{if } R_{semantic} < Z \\ 1 & \text{otherwise} \end{cases}, \quad (4.4)$$

where Z is a threshold parameter that determines how much semantic relevance a target image must have to be deemed relevant to the query, and thus the precision-recall curve can be plotted using the standard equations. Experimentally, there is little difference in the shape of the precision-recall curves with different values of Z , however, smaller values of Z result in much higher precision for all values of recall. For the experiments described in the next section, we have used a value of $Z = 0.5$, which implies that half or more of the annotation keywords in the query image must exist in the retrieved target image for it to be marked as relevant.

4.4 Results and Discussion

The results of the performance investigation are discussed as follows; firstly, the effect of the various parameters on the performance is discussed, and then overall precision-recall and semantic relevance results for the two data-sets are presented using the optimal parameters.

4.4.1 The Vocabulary

The choice of a good vocabulary is essential for achieving high retrieval performance. We investigate two parameters of the vocabulary below; the vocabulary size (the number of terms it contains), and the sensitivity of the vocabulary (how well a vocabulary works with different images to the ones it was trained on).

4.4.1.1 Vocabulary size

The size of the vocabulary is particularly important. Recall that the process of vector-quantising the descriptors to terms in the vocabulary is analogous to stemming real words. If the vocabulary is too large, the words will remain un-stemmed and unique.

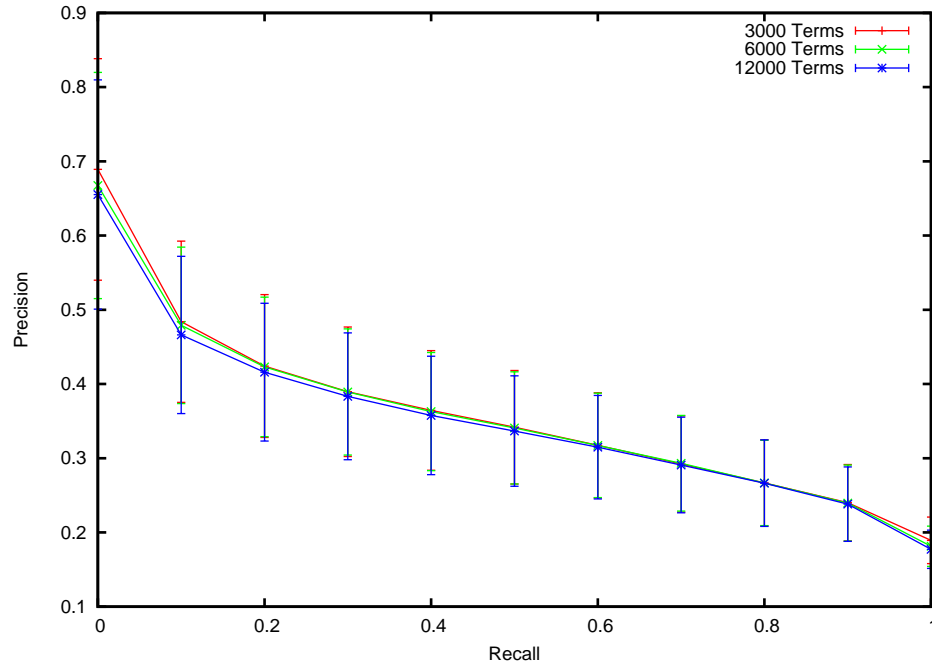


FIGURE 4.6: Precision-recall curves for different sizes of vocabulary using SIFT ‘visual’ terms with vector-space retrieval.

Number of Terms in Vocabulary	Semantic Relevance	
	Average rank-1	Average top 5
3000	0.50	0.44
6000	0.49	0.43
12000	0.47	0.41

TABLE 4.2: Semantic Relevance for different sizes of vocabulary using SIFT ‘visual’ terms with vector-space retrieval.

This would result in the vector-space being such that all documents are equally dissimilar, or orthogonal. Words with similar meaning would fail to be grouped together. If, on the other hand, the vocabulary is too small, words with differing meanings will be jumbled together. This will result in the documents all appearing similar (parallel) to each other in the vector-space.

Using the Washington data-set we generated vocabularies for 3000, 6000 and 12000 SIFT ‘visual’ terms. The retrieval performance using the vector-space can be assessed by comparing the precision-recall curves in Figure 4.6 and the semantic relevance’s in Table 4.2. The averaged precision of all of the vocabularies is very similar. For the remaining experiments we choose to use the 3000 term vocabulary because it is technically the best performing, although, because the results are so similar, this is hard to justify on this basis alone. However, we can better justify this choice because a smaller vocabulary gives a lower dimensionality vector-space which leads to much more efficient searching.

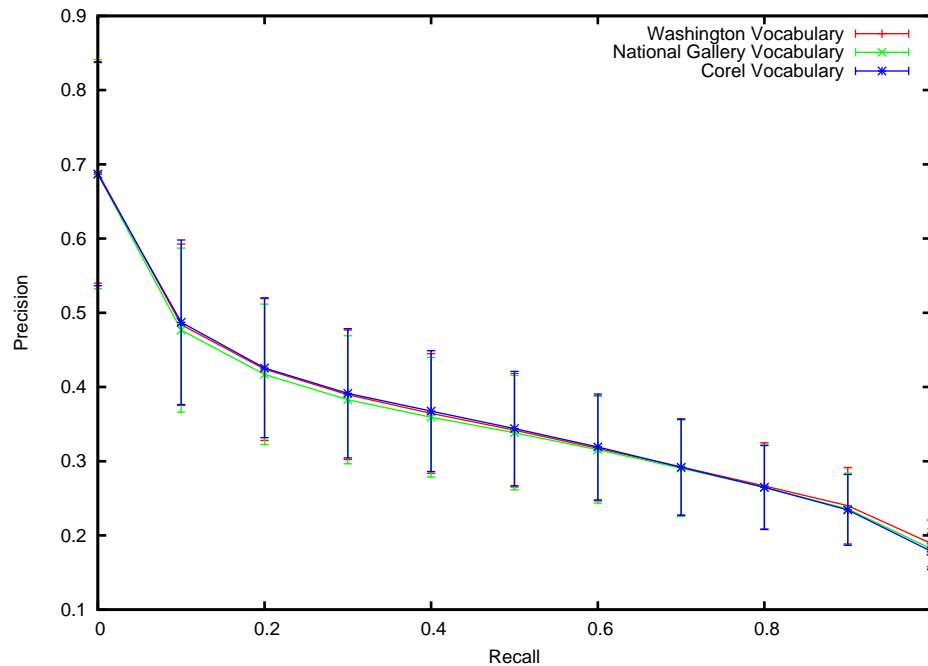


FIGURE 4.7: Precision-recall curves for three different 3000-term vocabularies using the Washington data-set with vector-space retrieval ($Z = 0.5$).

4.4.1.2 Sensitivity of retrieval with different vocabularies

It is interesting to study the sensitivity of retrieval with different vocabularies (vocabularies trained on different data) because creating a vocabulary using k -means is a computationally intensive process. For example, it took of the order of a few hours to create a 3000 term vocabulary using 100,000 samples of SIFT features, and of the order of days to create a 12000 term vocabulary. If it is possible to create a *universal* vocabulary that works well with all data-sets, then a lot of computational power and time can be saved.

In order to test how sensitive retrieval is to vocabularies trained on different training data, we generated three 3000-term vocabularies using SIFT keys from the Washington data-set, Corel data-set and National Gallery data-set. The National Gallery data-set is introduced in more detail in Chapter 5, but briefly, it consists of about 845 medium resolution (800-850 pixels on the longest dimension) scanned images of paintings from the National Gallery in London. Using these three vocabularies, we have calculated average precision-recall curves for vector-space retrieval as shown in Figure 4.7.

Figure 4.7 illustrates that there is virtually no difference in the average retrieval performance when using each of the three vocabularies. This implies that *universal* vocabularies are indeed possible. In order to verify this claim, we need to compare the vocabularies on a per query basis to ensure that there is little variation. In order to do this, we have plotted Relative R-Precision histograms between the Washington vocabulary and Corel vocabulary, and between the Washington and National Gallery vocabularies, as shown

if Figure 4.8. The R-Precision histograms show that on the whole the performance between the vocabularies is equivalent. Only on a few queries does the performance vary by a more significant amount.

4.4.2 Optimal k

The k value for LSI-based retrieval represents how many dimensions of the decomposed term-document matrix we believe are not attributed to noise. In practical retrieval scenarios with un-annotated data-sets, the value of k would have to be estimated empirically, based on some measure of perceived retrieval performance. However, in the case where we have an annotated data-set, such as the Washington data-set, it is possible to investigate the variation in a retrieval performance parameter, such as the average rank-1 semantic relevance over a range of k -values, and thus choose an optimal value to use for retrieval.

Figure 4.9 shows the variation of the rank-1 semantic relevance averaged over all queries with respect to the k value for LSI queries with different weightings and a 3000 term vocabulary. The figure shows that optimal retrieval appears to be at a k value of about 47 for each of the weightings.

4.4.3 Retrieval performance with the Washington data-set

In order to assess the performance of these retrieval techniques on the Washington data-set we performed a series of experiments to calculate precision-recall curves and semantic relevance. The experiments were performed using unweighted word occurrence vectors, in addition to vectors weighted using the tf-idf and log-entropy weightings described previously. The performance of the techniques in terms of their semantic relevance is summarised in Table 4.3. The results are compared with retrieval using a 64-bin grayscale histogram and ranking using the Euclidean distance. The grayscale histogram was chosen as it represents the lowest-denomination invariant image content descriptor that doesn't use colour information, like the SIFT descriptors.

Table 4.3 shows that LSI-based retrieval (with $k = 47$) outperforms the vector-space method by a small margin, and both methods are much better than retrieval through global grayscale histograms, and certainly much better than random retrieval. The best weighting for the LSI technique appears to be log-entropy, and the vector-space model works best without any weighting applied. Figure 4.10 shows precision-recall curves for the experiments. As hinted in Table 4.3, the log-entropy weighted LSI approach achieves the highest precision. However, the log-entropy LSI curve degrades much more rapidly than the unweighted vector-space curve. The unweighted vector-space curve follows the same shape as the grey-level histogram curve, albeit with a 10% higher precision across

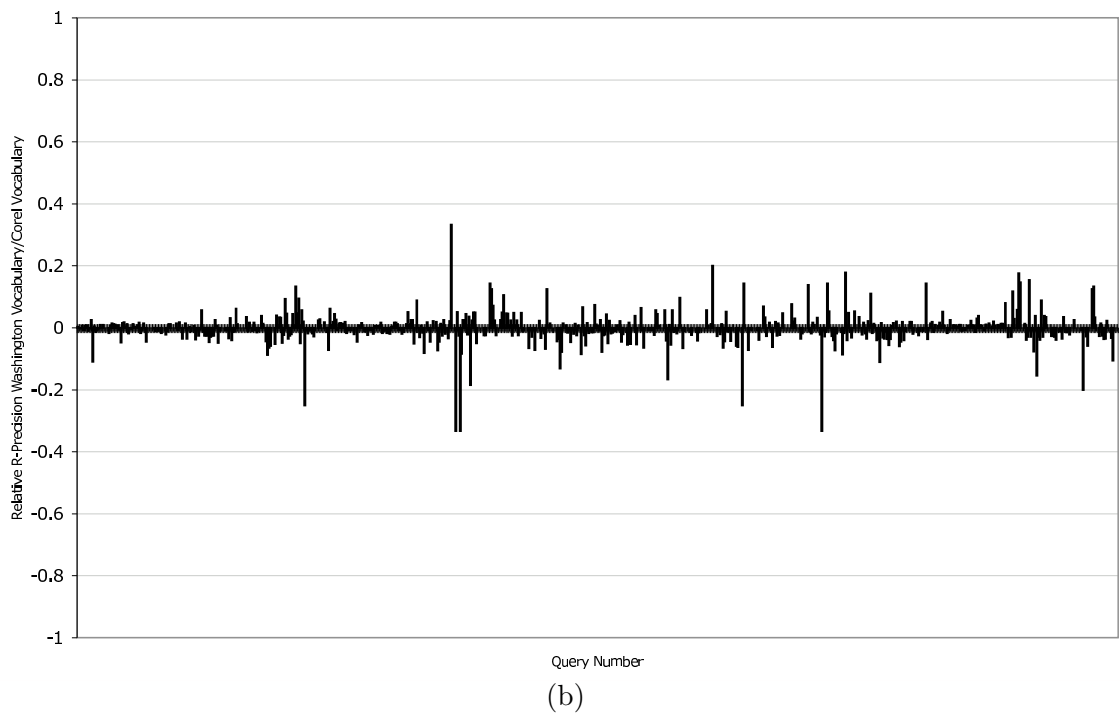
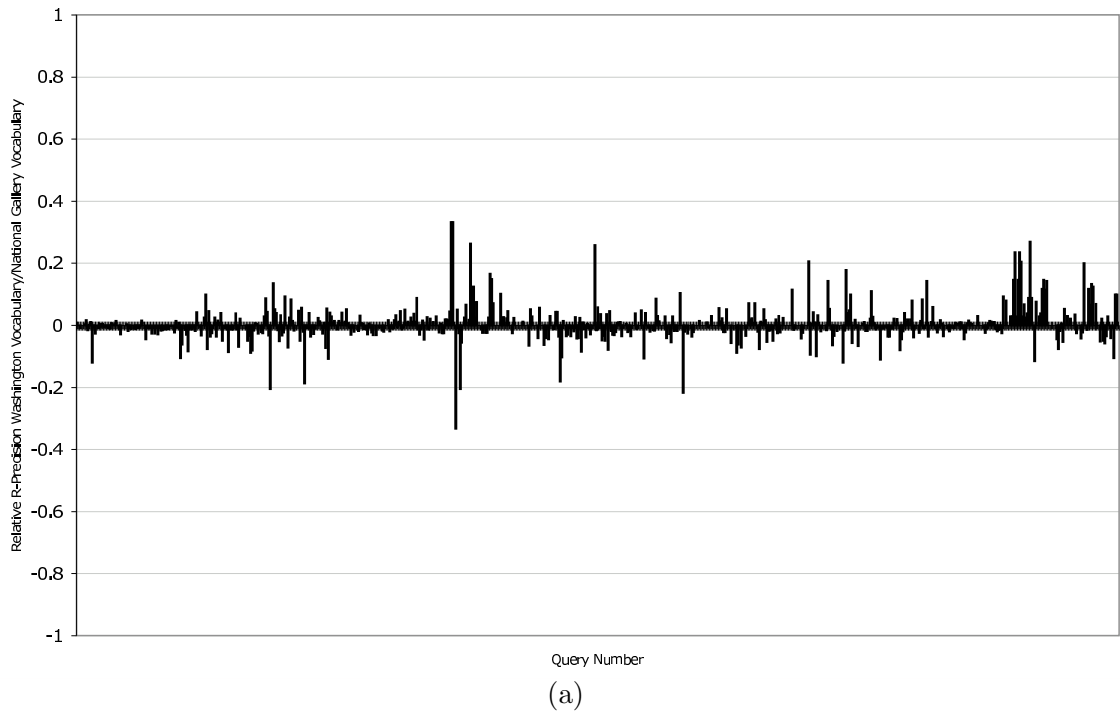


FIGURE 4.8: Relative R-Precision histograms showing the relative performance of retrieval using different vocabularies. (a) Shows the Washington vocabulary versus the National Gallery vocabulary, and (b) show the Washington vocabulary against the Corel vocabulary. $Z = 0.5$ in both cases.

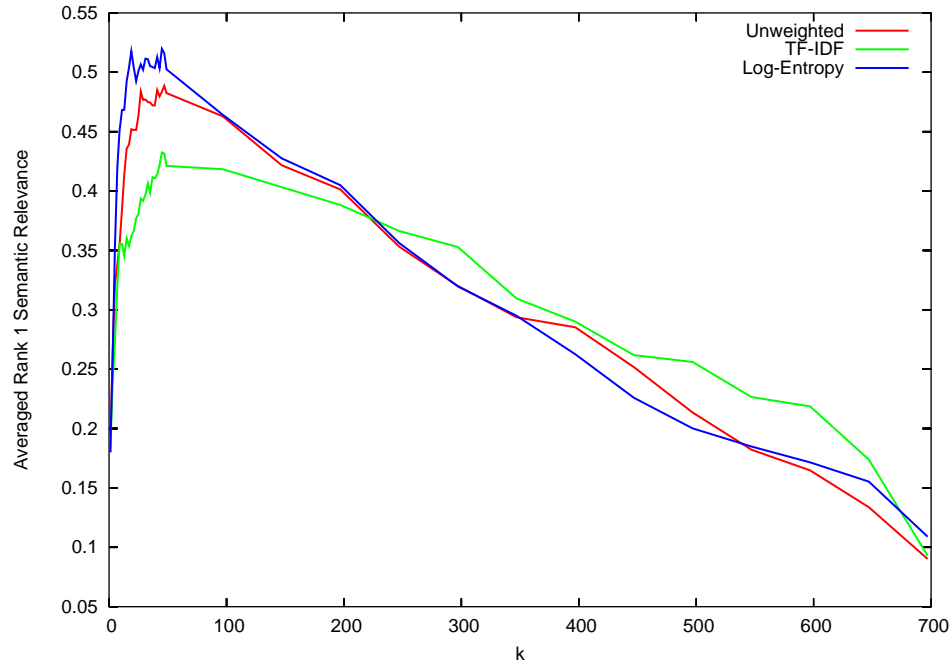


FIGURE 4.9: Effect of varying k with respect to retrieval performance for LSI based retrieval.

Method	Weighting	k	Rank 1 Semantic Relevance	Averaged Top 5 Semantic Relevance
LSI	Unweighted	47	0.53	0.44
	TF-IDF	47	0.48	0.41
	Log-Entropy	47	0.56	0.46
Vector Space	Unweighted	N/A	0.50	0.44
	TF-IDF	N/A	0.48	0.42
	Log-Entropy	N/A	0.45	0.40
64 bin Grayscale Histogram	N/A	N/A	0.41	0.35
Random Retrieval	N/A	N/A	0.14	0.14

TABLE 4.3: Summary of average semantic relevance values for retrieval with the Washington data-set using SIFT-based ‘visual’ terms together with Vector-Space and LSI techniques.

most values of recall. These results indicate that the LSI approach can give us better results than the vector-space approach when we are only interested in looking at the top few similar images (i.e. recall is low). If more images are required, then the vector-space model out-performs the LSI approach.

It is also interesting to investigate the effect that different image feature morphologies have on retrieval performance. In order to do this we have investigated the performance of the two text retrieval techniques using ‘visual’ terms from our colour descriptor. In addition, we used the techniques with a combined image description formed by appending the SIFT word occurrence vector with the colour word occurrence vector for each image respectively. As before, the k value for LSI was optimised and found to be optimal at

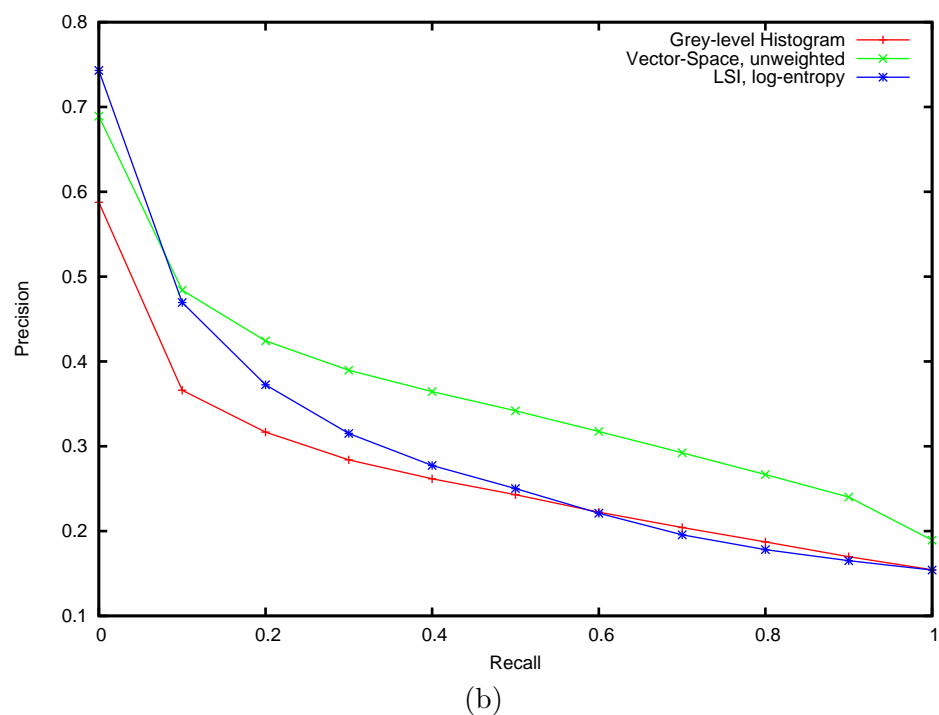
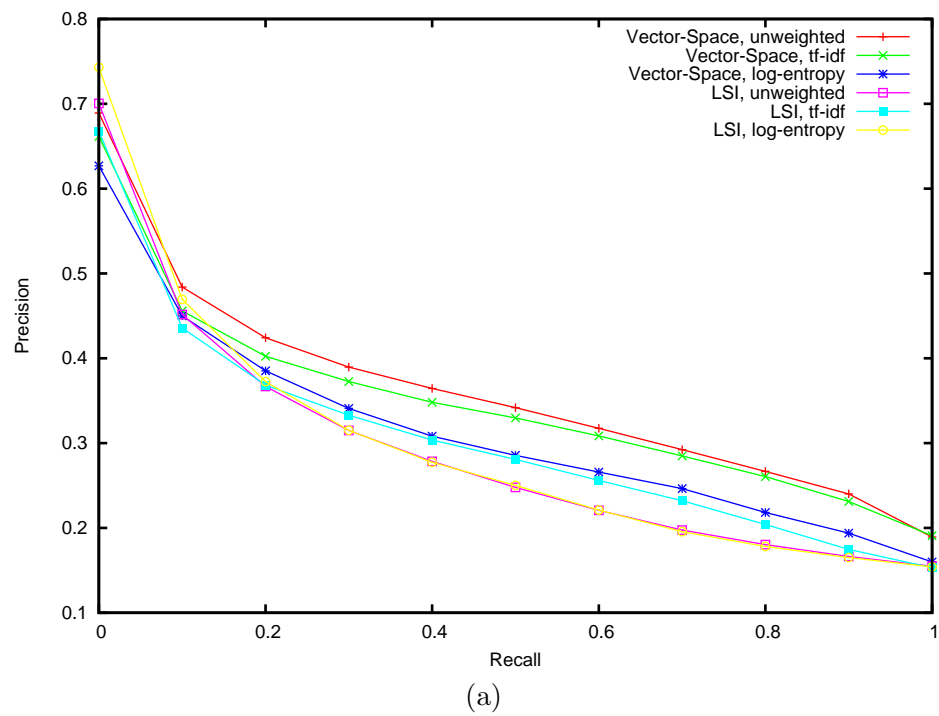


FIGURE 4.10: (a) Average precision-recall curves for the Washington data-set using different weighting schemes with SIFT ‘visual’ terms. (b) Average precision-recall curves for the Washington data-set with the best performing weightings using SIFT ‘visual’ terms and using grayscale histogram retrieval.

a value of 10 for the colour terms and 48 for the combined terms. Precision-recall plots showing the results of these experiments are shown in Figures 4.11 and 4.12 together with the result of retrieval using $4 \times 4 \times 4$ -bin RGB histograms with Euclidean ranking.

Figure 4.11 shows that the performance of both the LSI and vector-space approaches when coupled with the colour ‘visual’ terms is very similar to the performance when retrieving using RGB histograms. This indicates that the ‘visual’ term occurrence vector is approximating the colour distribution within the image. The effect of combining term occurrence vectors as shown in Figure 4.12 improves retrieval for some of the weighting schemes, most notably vector-space retrieval with log-entropy weighting. However, none of the combined term vector perform as well as with the SIFT term only vectors. This gives us an indication that the semantics of the Washington data-set, in the form of the keyword annotations are not well modelled by colour information. This is not necessarily surprising because not many of the keywords used to annotate the data-set have specific, unique colours associated with them. This issue is discussed again in more detail in Chapter 6.

4.4.4 Retrieval Performance with the Corel Data-set

The Corel data-set demonstrates some problems with the vector-space image description and retrieval approach described. Figure 4.13 shows the averaged precision-recall curves for a retrieval experiment using the same methodology as above. Curves are shown for global RGB- and mono-histogram retrieval, in addition to retrieval with the SIFT ‘visual’ terms (3000-term vocabulary) using the un-weighted vector-space and LSI approaches. The curve for retrieval using the vector-space model together with colour ‘visual’ terms is also shown. Different weightings are not shown as they had negligible effect on the performance.

The curves in Figure 4.13 show that retrieval using ‘visual’ word representations of the Corel images is not much better than using the grayscale histogram retrieval method; the techniques actually perform slightly worse at a recall of less than about 0.1. This is in contrast to the results from experimentation with the Washington data-set. All of the curves show the same general trend, with a high initial precision with a sharp drop-off to a relatively flat curve after a recall value of about 0.3 has been attained. Overall, the global RGB histogram gives the best retrieval.

The shape and general trend of the precision-recall curves is related to the image content and the keywords used to annotate each of the images. In the past, the Corel collection has been criticised for being a particularly easy collection from a retrieval point-of-view. Müller et al. (2002) discussed the Corel image collection in detail, and showed how different data-sets could be created from subsets of the collection in such a way as to improve the apparent performance of a retrieval system. Some of the images in the

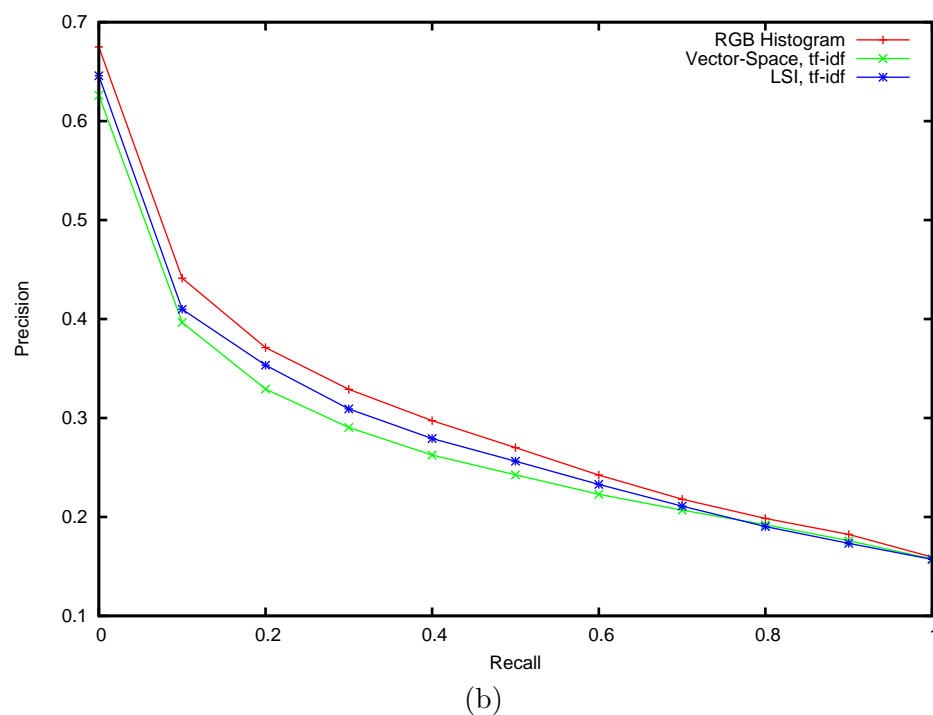
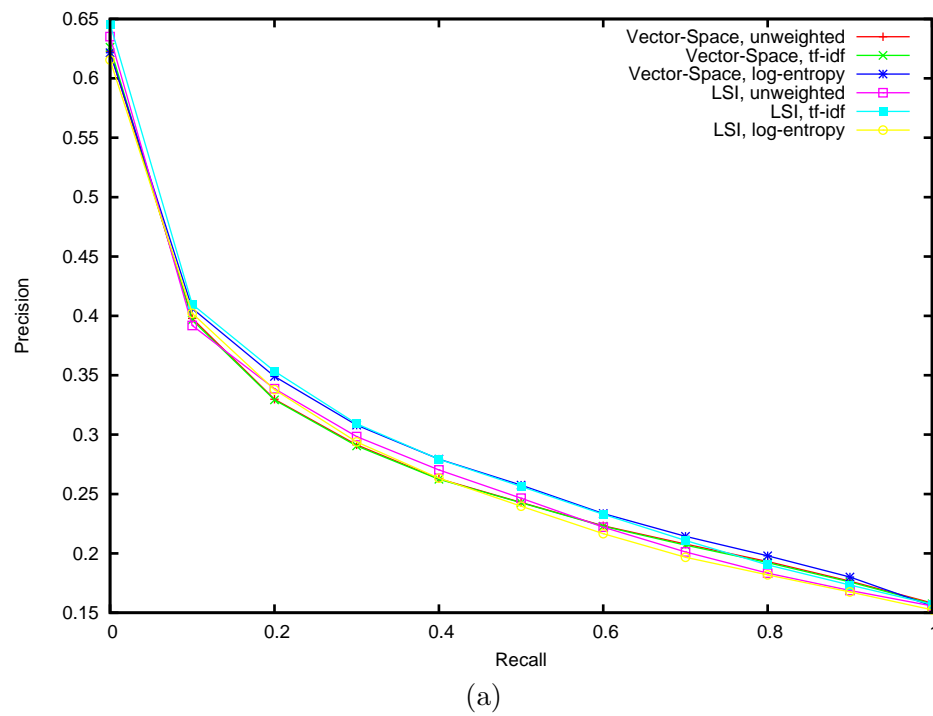


FIGURE 4.11: (a) Average precision-recall curves for the Washington data-set using different weighting schemes with colour ‘visual’ terms. (b) Average precision-recall curves for the Washington data-set with the best performing weightings using colour ‘visual’ terms and using RGB and grayscale histogram retrieval.

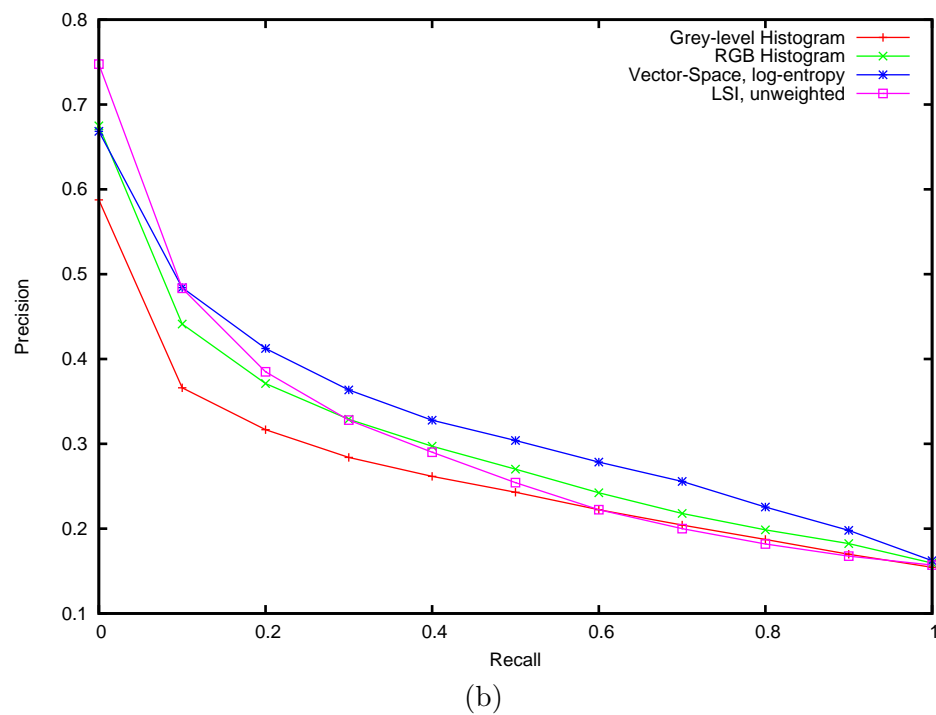
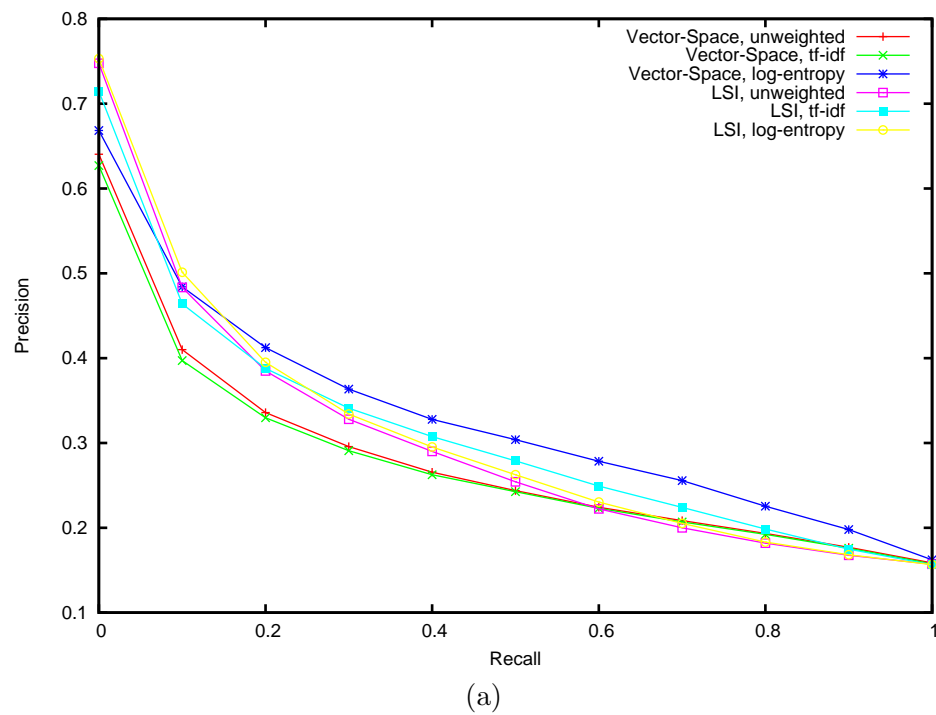


FIGURE 4.12: (a) Average precision-recall curves for the Washington data-set using different weighting schemes with combined SIFT and colour ‘visual’ terms. (b) Average precision-recall curves for the Washington data-set with the best performing weightings using combined SIFT and colour ‘visual’ terms and using RGB and grayscale histogram retrieval.

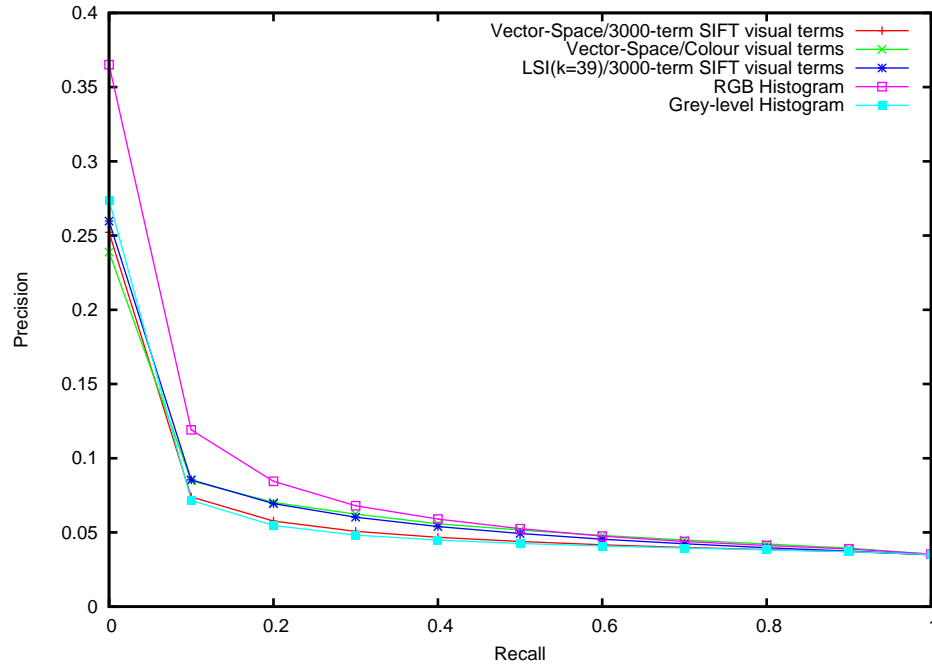


FIGURE 4.13: Average precision-recall curves for the Corel data-set.

Corel collection are particularly easy to find using global colour information, such as those showing *sunsets*. However, these images are rarely labelled as containing just the *sun*, but they may also contain other labels, such as *mountain* and *city*. In terms of semantic relevance, this can cause a problem; given a query image of a *sunset*, other *sunset* images are easy to find, leading to high precision. However, given a query image predominantly showing a *city*, an image of a *sunset* and *city* is unlikely to be ranked very high, leading to low precision at high recall. The salient region approach to modelling the image content was designed to avoid this problem, however, it appears to fail in the case of the Corel data-set.

In the Washington data-set, the retrieval performance using the vector-space model with colour ‘visual’ terms was seen to be fairly similar to the RGB histogram retrieval. This was expected, because the distributions of (uni-modal) dominant local colour from all of the salient regions should fairly well approximate the global colour distribution. However, in the Corel data-set this is not really the case.

The problem lies in the the size of the images; because we only have access to thumbnail sized images, the lack of any relatively high frequencies in the image makes it very difficult to extract many difference-of-Gaussian salient regions. On average, most of the images only appear to have between 10 and 20 salient regions. This is in stark contrast to the thousands of regions detected in each of the Washington images. The lack of regions means that the term-occurrence vectors representing each image are incredibly sparse, leading to a poor feature space.

Although not explored here, it may be possible to avoid this problem by combining the output of multiple region detectors which will give a much richer image description. However, it is quite possible that even that approach will fail on images of this resolution.

4.4.5 Computational Performance

Hitherto, we have not discussed the relative performance of the two approaches in terms of their computational complexity. Whilst the LSI algorithm trades higher precision at low recall for much lower precision at higher recall when compared to the vector-space algorithm it does have an advantage in that it reduces the dimensionality of the search space dramatically. Searching a 47 or so dimensional space versus searching in 3000 dimensions gives a massive speed advantage due to the brute-force methods used, although this is offset somewhat by the time taken to calculate the matrix decomposition. If, however, the data-set is static, this is a one-off cost as it only needs to be done once. It is beyond the scope of this thesis to investigate this in more detail, however it would be interesting to see how an inverted-index approach to indexing the vector-space performs in terms of its computational speed when searching the reduced LSI space. The efficiency of a inverted index approach would be somewhat related to the density of the term occurrence vectors.

4.5 Conclusions

This section has presented a way to link methods from the information retrieval community with image description through salient regions to form powerful image retrieval techniques. We have shown how local descriptors from salient regions can be quantised into ‘visual’ terms and these terms used as a basis for indexing through the vector-space and Latent Semantic Indexing retrieval models.

Evaluation of the two techniques on the Washington data-set has shown that with well-chosen parameters, the LSI technique exhibits a slightly better performance than the vector-space technique at low values of recall, but performs worse as recall increases. Both techniques vastly outperform retrieval by global grayscale histogram matching.

Experiments with the thumbnail images from the Corel data-set showed less promising results, but subsequent investigation has shown this to be due to the lack of high-frequency information within the images from which to select salient regions. This lack of salient regions causes the ‘visual’ term-occurrence vector-space to be poorly defined.

4.6 Summary

This chapter has described how image retrieval can be performed using image representations from local descriptors of salient regions, as described in the previous chapter. The main contribution of the chapter has been to investigate how techniques from the text retrieval community can be exploited for use with these image descriptions. The chapter has also introduced a technique for assessing the content-based retrieval performance of annotated image collections in query-by-image-content type tasks. The chapter concluded with a discussion about the relative performance of the two text retrieval approaches investigated.

Chapter 5

Query By Mobile Device

“...when you have eliminated the impossible, whatever remains, however improbable, must be the truth.”

SIR ARTHUR CONAN DOYLE, SHERLOCK HOLMES

This chapter aims to demonstrate the robustness of the vector-space retrieval approach discussed in the previous chapter. Image descriptions from SIFT features of difference-of-Gaussian salient regions applied to the vector-space model are shown to outperform other retrieval approaches in the retrieval scenario described here.

The chapter introduces a new paradigm for content-based image retrieval, in which a mobile device is used to capture the query image and display the results. The system consists of a client-server architecture in which query images are captured on a mobile device and then transferred to a server for further processing. The server then returns the results of the query to the mobile device. There are a number of possible user-scenarios for the use of such a device. These scenarios generally fall into two categories, depending on what kind of query result the system would be expected to provide.

The first category is very much like previous research on the “physical hyper-link” carried out at HP labs (Barton and Kindberg, 2001), where a user can ‘click’ on real world objects as if they were a hyper-link, using a mobile device as the interface. In this case, the objective of the system is to find an *exact* representation of the query image in the database and to return metadata corresponding to the object represented in the query image. For example, consider using the device in a museum or art gallery. The device could be pointed at various exhibits or paintings and would return metadata about that particular object. Another possible example would be in a bookshop. In this case the device could be pointed at a book cover, and the returned metadata could be, for example, reviews of that particular book.

The second category is much more like classical content-based image retrieval. In this case, the objective is not necessarily to find an exact match, but rather to find a ranked set of *similar* images - either visually similar (e.g. in terms of colour) or similar in terms of the semantics of the content.

This chapter examines the first category in detail, although the retrieval algorithms presented are equally applicable to the second category. The chapter is split into several sections. The first section discusses some of the problems and requirements with retrieval from a mobile device. The second section shows how the vector-space retrieval model from the previous chapter has been augmented to fulfil the requirements. The third section shows how the retrieval model has been implemented in a client-server architecture. The fourth section illustrates some results of our system in a mock museum scenario. Finally, the last section provides an executive summary of the chapter.

5.1 Requirements

The aim of the system described in this chapter was foremost to demonstrate the power of the retrieval approach described in Chapter 4. The scope of the system was limited to cover image retrieval of paintings from a mobile device within an art gallery. The idea was that the mobile device could be used to query a painting hanging on the wall, and that the device would show metadata about the artwork, perhaps in the form of a web-page. Figures 5.1 and 5.2 illustrate the idea with montages of screen-shots from the second of our demonstration implementations.

It was decided that the system should be able to work with current mobile hardware technology. State-of-the-art mobile devices, such as camera phones, have built in cameras for image capture, and the ability to connect to the internet through systems such as GPRS. What most current mobile devices lack, however, is computational power, for example most current devices are unable to natively perform floating-point maths. These constraints meant that the system had to be designed in a client-server fashion, with the mobile client handing off the majority of processing to the server.

Constraining the system to work only in an art gallery scenario with paintings simplifies the retrieval somewhat. The fine-art paintings we dealt with were flat surfaces, this meant that the retrieval algorithm would only have to deal with planar homographic transformations between the query image and the images in the database (there are some other geometric imaging issues such as warping due to the camera lens, but these can be removed through calibration if necessary). The difference-of-Gaussian salient regions described in Chapter 3 were shown to be quite robust to this kind of transform; certainly within the limits we envisaged the query images to be captured from.



FIGURE 5.1: Montage showing a screen-shot from the software demonstrator in capture mode and the artwork being captured. Images Copyright © 2005, National Gallery, London, All rights reserved.



FIGURE 5.2: Montage showing various parts of the metadata shown to a user by the software demonstrator as they scroll through it. Images and Metadata Copyright © 2005, National Gallery, London, All rights reserved.

5.2 Approach

The retrieval approach is taken from the work described in the second half of Chapter 4; Images are indexed using a vector-space formed from ‘visual’ term-occurrence vectors. The ‘visual’ terms are created from quantised SIFT descriptors from salient regions within each image. This representation allows images from a database to be ranked according to similarity to the query image.

5.2.1 Geometry-based Re-Ranking

Due to the way the indexing scheme works the top ranking matching image may not actually be a representation of the query image. This is due in part to the imaging conditions, but also to the fact that the query image is likely to be either a sub-image or super-image of the matching representation in the database. In order to find the actual matching image, we re-rank the top N results based on the geometric consistency of the salient regions. This is akin to text-search methodologies where the rank of the document is increased if the query terms occur with similar positional relations to each other in both the query and document.

Because the aim of the system is to recognise planar objects, we model the geometric consistency of the salient regions as a planar homography. In order to perform the re-ranking, we test each of the top N ranked images' salient regions for a consistent homography between the query image's salient regions using the RANSAC algorithm to robustly ascertain whether a consistent homography exists (Vincent and Laganière, 2001). An alternative approach, not explored here, would be to use a geometric hashing approach (Schwartz and Sharir, 1987; Wolfson and Rigoutsos, 1997), or by clustering features in *pose-space* using a Hough transform (Lowe, 2004).

5.2.2 Summary

In summary, we have presented an extension to the image retrieval methodology described in Chapter 4 with a two-stage re-ranking procedure. The algorithm transforms the query image into a vector-space based on the frequencies of 'visual' words within the image. The 'visual' words are created in such a way as to be invariant to a range of transformations, including changes in homography, intensity changes and imaging noise. The first stage ranking procedure uses the cosine similarity of weighted 'visual' word frequency vectors to rank the images in the database. The second stage re-ranks the top N results based on the geometric consistency between the salient regions of the query and N results. The outcome is that the highest ranked image should correspond to the query. The overall retrieval process is illustrated in Figure 5.3.

5.3 Client-Server Implementation and Technology

In order to develop our mobile architecture for retrieval, a test-bed has been constructed from commercially available equipment, and using open standards for data transfer. The first implementation of the system consisted of a mobile device with a camera (an HP h5550 iPAQ Pocket PC and Lifeview FlyCAM SD) acting as a mobile client, and a PC acting as a server. The mobile client is connected to the Internet through a wireless connection (either Bluetooth or 802.11b). The server machine hosts a web service to which the client can connect and transmit JPEG compressed query images. XML remote procedure calls (XML-RPC) are used to provide the interface to the server. The server processes the queries it receives and returns the result to the client. Figure 5.4 illustrates the topology of the system. The second implementation of the system consisted of a software demonstrator that captured images through a webcam and connected to the server as before. The second demonstrator aimed to illustrate how the system would look if it worked on a mobile phone (Figures 5.1 and 5.2).

Figure 5.5 illustrates the use of the device in an art gallery scenario. The server has been configured to return a web-page with information corresponding to the database image that most closely matches the query. The web-page is then displayed on the client.

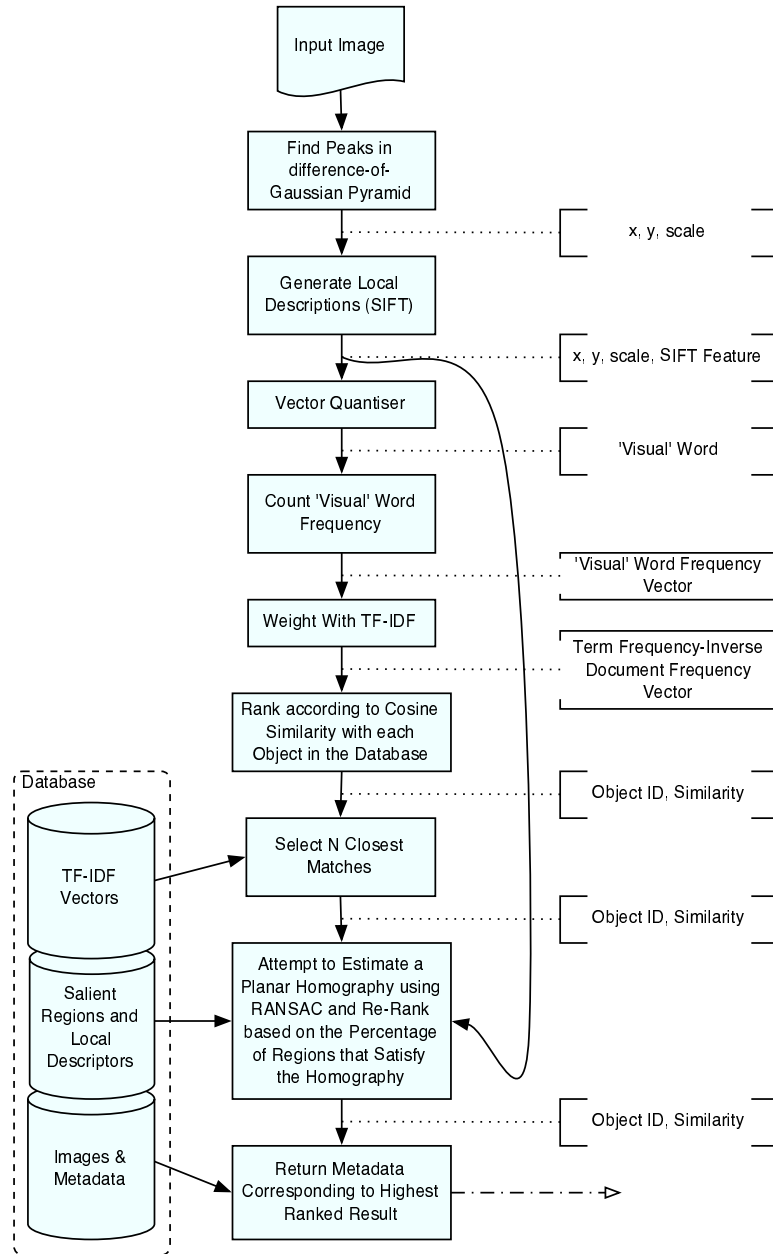


FIGURE 5.3: Overview of our content-based image retrieval technique.

5.4 Retrieval Performance

The performance of the retrieval algorithm was evaluated by testing 200 randomly selected images captured using the mobile device and looking at the rank of the matching image in the returned set. Obviously, the ideal scenario is that the matching image is always returned in the highest ranking (rank 0) position. The image database consisted of over 850 images from the National Gallery image collection. A number of sample query images are shown in Figure 5.6.

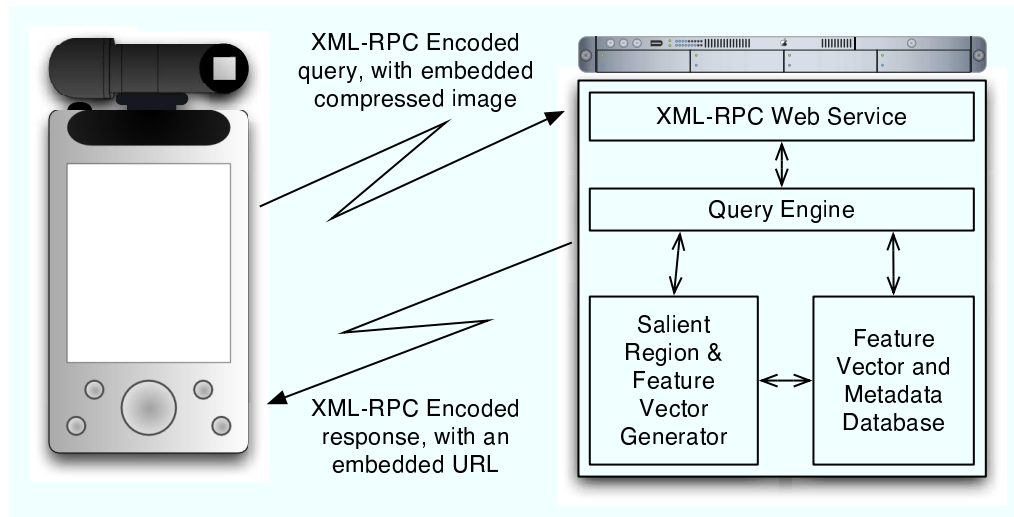


FIGURE 5.4: An overview of the mobile image retrieval system.

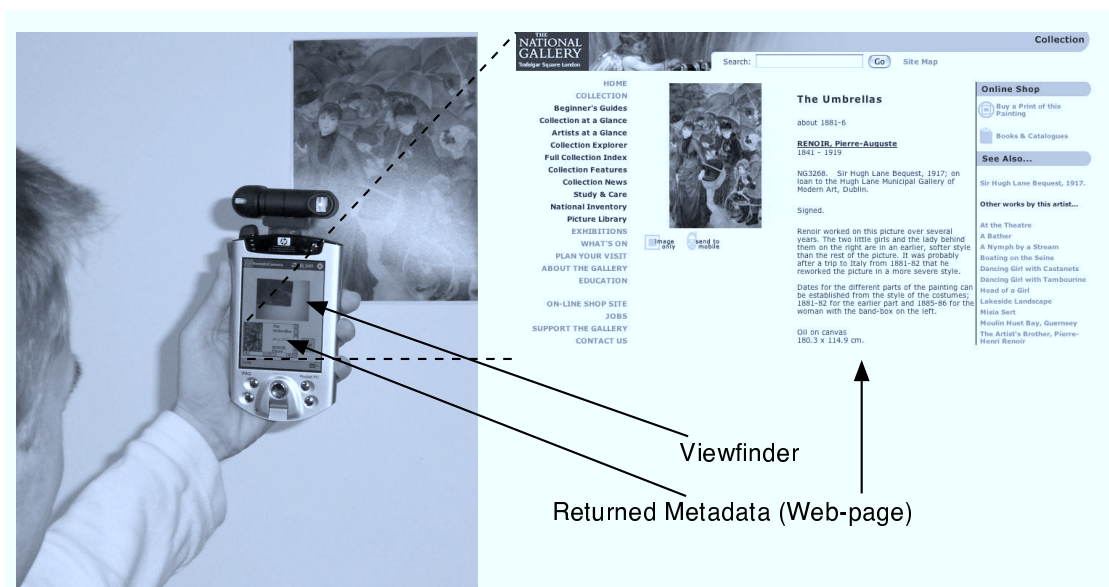


FIGURE 5.5: The system in use in a mock art gallery scenario. Images Copyright © 2005, National Gallery, London, All rights reserved.

Figure 5.7 illustrates the effect of querying the database with a number of different retrieval algorithms, including the Colour Coherence Vector (CCV) algorithm (Pass et al., 1996), RGB Colour Histogram matching, Grey-level Histogram matching, Pyramid-structured Wavelet Transform (PWT) algorithm (Fauzi and Lewis, 2002), and the vector-space retrieval algorithm detailed in the previous sections, without the second-stage re-ranking. The graph shows that the vector-space retrieval algorithm performs dramatically better than the other algorithms; in fact, the performance of the other algorithms is little better than randomly choosing an image from the database. Just under 35% of matching images using the vector-space algorithm were found in the highest ranking position, and the percentage of matched images drops off rapidly as rank increases.



FIGURE 5.6: Example query images captured by the mobile device for testing the performance. Images Copyright © 2005, National Gallery, London, All rights reserved.

The effect of the second-stage re-ranking was also investigated. The purpose of the two-stage re-ranking approach is to reduce computational load. The first retrieval stage identifies possible matches, and the second-stage verifies the actual match. If the second stage re-ranking were performed on all the images in the database, the probability of identifying a correct match is extremely high, but the computational load would be massive and the need for the first-stage retrieval would be negated. By considering only the top N ranking matches from the first-stage in the second-stage, computational load is dramatically reduced at the expense of retrieval performance. Figure 5.8 illustrates

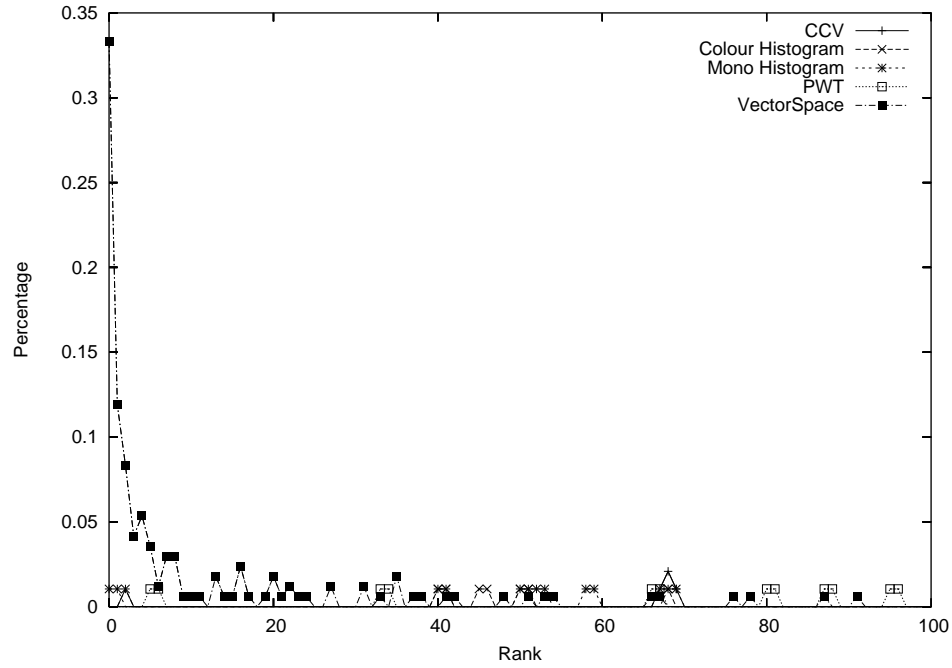


FIGURE 5.7: Plot of the rank of the matching image for a number of different retrieval algorithms.

the effect of changing N versus the rate of correct retrieval, where correct retrieval is defined as the image matching the query being in the highest ranking position after the second-stage re-ranking. The graph shows that a first-place recognition rate in excess of 80% can be achieved by performing the geometry based re-ranking procedure on the top 20 ranked matches from the first-stage retrieval.

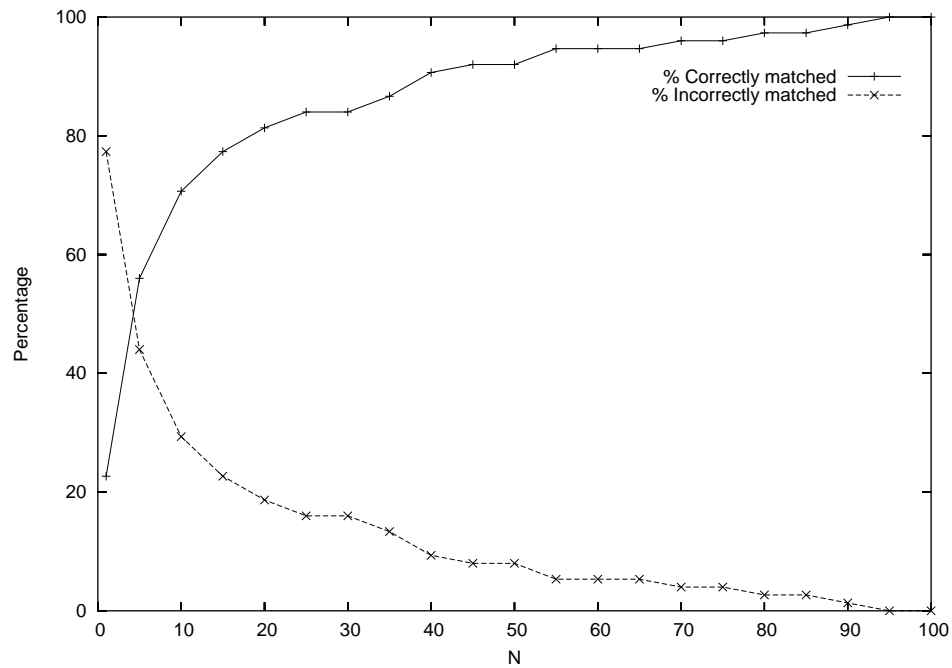


FIGURE 5.8: Retrieval rate versus N , the number of images considered for second-stage geometry based re-ranking.

5.4.1 Discussion

The results presented above were found using a naive set of parameters for things such as the number of ‘visual’ words in the vocabulary. It is possible that by tuning the parameters, the retrieval performance could be further improved. The vector quantiser used for the experiments was certainly non-optimal for the test image data-set, and no investigation into the optimal number of ‘visual’ words in the vocabulary was performed. Performance could also possibly be improved by pre-processing the query images to remove the radial lens distortion the camera exhibits and also by normalising the images. However, despite these non-optimised parameters, the results show that the two-stage retrieval algorithm performs well when presented with query images of low quality, such as those from a mobile device.

5.5 Summary

This chapter presented an investigation into the use of a mobile device as a novel interface to a content-based image retrieval system. The chapter presented a novel methodology for performing content-based image retrieval and object recognition from query images that have been degraded by noise and subjected to transformations through the imaging system. The methodology used techniques inspired from the information retrieval community in order to aid efficient indexing and retrieval. In particular, a vector-space model was used in the efficient indexing of each image, and a two-stage pruning/ranking procedure was used to determine the correct matching image. The retrieval algorithm was shown to outperform a number of existing algorithms when used with query images from the mobile device.

Chapter 6

Auto-Annotation and Advanced Retrieval

“There is nothing worse than a sharp image of a fuzzy concept.”

ANSEL ADAMS

Searching an image collection can be made intuitive when adequate annotations are available. The keyword terms used for annotation are inherently semantic. By performing text query searches using standard techniques against the keyword terms, images can be found in a manner that will satisfy many users. Of course, this technique can also be combined with visual content search techniques to give the user much more control over the search.

The standard approach to enabling keyword searching of image databases has been to attempt to apply automatic annotation to automatically generate the keywords for un-annotated images. Previous approaches to automatic image annotation have tended to use region-based image descriptions, typically generated by automatic segmentation or through fixed, usually rectangular, shapes. Rectangular regions are a poor choice for image description because they are not robust to a variety of transformations, such as image rotation. The segmentation approach has a large problem — that of how to perform the segmentation. Over the years many techniques for performing image segmentation have been suggested, although none can really solve the problem of linking the segmented region to the actual object that is being described. Indeed, this shows that the non-naive segmentation problem is not just a bottom-up image processing problem, but also a top-down problem that requires prior knowledge of the true object, before it can be successfully segmented.

This chapter discusses two approaches to enabling keyword searching of un-annotated image databases using techniques developed in the previous chapters. The first approach

uses the vector-space representation of the local descriptors of salient regions to describe the image in an invariant manner, together with a method of semantic propagation to generate the correct annotations for the image.

The second approach does not actually explicitly provide annotations for un-annotated images, but instead uses a generalisation of a linear-algebraic technique known as Cross-Language Latent Semantic Indexing in order to create a semantic space of images and terms. This semantic space can be queried using keyword terms, and aims to return images related to that term.

6.1 Auto-annotation using Semantic Propagation

This section presents our model of automatic annotation based on the propagation of semantics. The premise behind the model is intuitive; images that are *visually* similar often have similar meaning or semantics.

Using the vector-space and Latent Semantic Indexing techniques together with ‘visual’ words as discussed in Chapter 4, we have all the tools needed to compare and rank documents based on their visual content. By creating a collection or corpus of pre-annotated images, it should be possible to label unannotated images by looking for similar annotated ones. In our preliminary model of annotation, we just apply, or propagate the labels from the closest M matching images to the unannotated query image.

The remainder of this section is devoted to describing an investigation into the plausibility and performance of this technique for auto-annotation. Results using only the SIFT visual terms on the Washington data-set (Hare and Lewis, 2005c) are presented. Section 6.2 compares this technique against a different technique for image retrieval based on keywords.

6.1.1 Preliminary Results

6.1.1.1 Image Dataset

The 697 annotated images from the Washington data-set were used for the preliminary investigation. We processed the annotations to correct mistakes and fold together terms by merging plurals into singular form (i.e. “trees” became “tree”). The original 287 keywords became 170 terms with these modifications. The average number of keywords per image is 4.8. The empirical keyword distribution across the dataset is shown in Figure 6.1. For experimentation, the dataset was randomly split into two parts, with one part used for training, and one part used for testing.

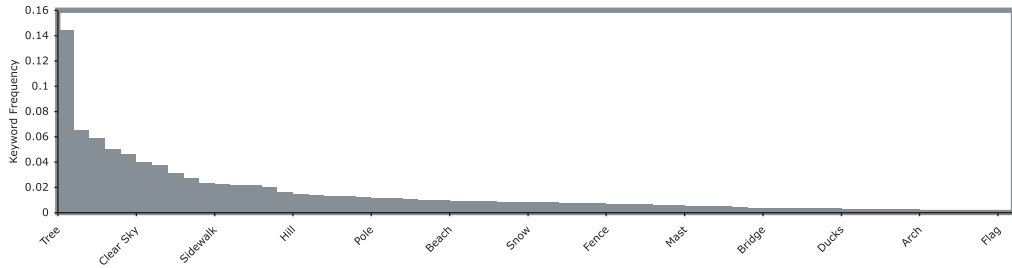


FIGURE 6.1: Plot of empirical keyword distribution in the dataset

6.1.1.2 Performance Evaluation

Many different measures could be chosen for evaluating the performance of an auto-annotation algorithm, but a number of factors need to be accounted for when choosing a measure. Firstly, the statistics of the vocabulary have to be taken into account. Figure 6.1 shows the empirical distribution of keywords in the dataset. Because words like ‘Tree’ occur more often, they are much safer guesses when determining annotations. An auto-annotation technique should therefore perform better than a technique that pseudo-randomly applies labels based on the empirical distribution of keywords in the training set.

Secondly, the training dataset itself might not contain *correct* keywords for some of its images. For comparative purposes, this is not a problem because all of the algorithms have to deal with the same data, however, in an absolute sense, the reported performance is likely to be overly pessimistic.

Thirdly, the performance measure needs to account for the number of incorrect words. An ideal auto-annotation system should choose the correct number of keywords required to describe the image content. Barnard et al. (2003), suggest the use of the *normalised score* measure, E_{NS} :

$$E_{NS}^{(model)} = \frac{r}{n} - \frac{w}{N - n}, \quad (6.1)$$

where r is the number of correctly predicted words, n is the actual number of keywords in the query image, w is the number of wrongly predicted words, and N denotes the number of words in the vocabulary. The score gives a value of 1 if the image is annotated exactly correctly, a value of 0 for predicting both everything or nothing, and -1 if the exact complement of the actual word set is predicted. The use of the normalised score is not without problems however. If we are to believe that the measure used should choose the correct number of keywords, then the normalised score is not a good measure as it does not sufficiently weight incorrect guesses. It can be seen from the normalised score equation that if the vocabulary is very large (large N) and n is modest, a significant number of wrong words (w) can be assigned without significantly affecting the score. For example, Monay and Gatica-Perez (2003) report that in their test database, with an average of 18.5 keywords per image, the normalised score is maximised when their annotation algorithms return about 40 keywords per image. This implies that even if

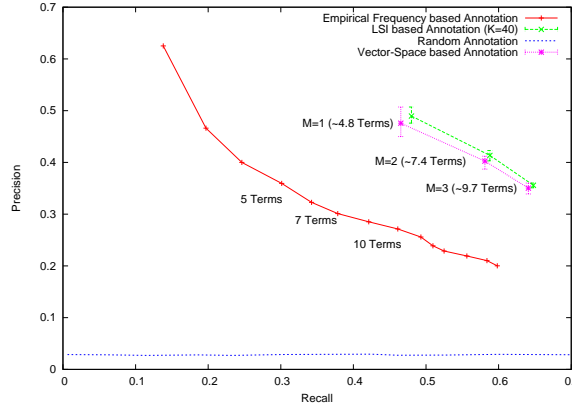


FIGURE 6.2: Precision-Recall curves for each of the auto-annotation methods. Error bars show range of precision over repeated runs

the annotation algorithm is selecting all of the correct labels, it is selecting even more incorrect ones, thus making for very noisy annotations.

In order to address this problem, we have chosen to use precision and recall as our measures for evaluation, although we do also include the normalised score for comparison. Using the same terminology as above, precision and recall are defined as:

$$Recall = \frac{r}{n} \quad (6.2)$$

$$Precision = \frac{r}{r + w} \quad (6.3)$$

The interpretation of the precision and recall measures for evaluation of auto-annotation are a little different from the evaluation of retrieval systems. In retrieval, the aim is to get a high precision for all values of recall. However in annotation, the aim is to get both high precision (high proportion of correctly guessed labels to the number guessed) and high recall (high overall proportion of correct labels).

6.1.1.3 Experimental Results

A number of experiments were performed to ascertain the performance of the two annotation methods and also to provide comparison of their performance against annotation using randomly selected labels, and labels selected based on the empirical frequency distribution in Figure 6.1. The experiments were performed using a randomly selected 50 : 50 mix of images from the dataset to provide a set of training images and a set of query images. The number of visual terms was set to 3000 (Hare and Lewis, 2005b). The word-occurrence vectors for both the vector-space and LSI models were unweighted. The optimal number of dimensions of the semantic space, K , for the LSI model was found to be about 40 with respect to maximising the precision, recall, and normalised score.

Method	M	Number of Words	Precision	Recall	E_{NS}
Vector-Space	1	~ 4.8	0.476	0.465	0.450
	2	~ 7.42	0.402	0.581	0.554
	3	~ 9.70	0.350	0.641	0.602
LSI (K=40)	1	~ 4.8	0.490	0.480	0.466
	2	~ 7.42	0.414	0.588	0.561
	3	~ 9.70	0.356	0.648	0.609
Empirical	-	5	0.329	0.343	0.323
	-	7	0.288	0.425	0.394
	-	9	0.241	0.509	0.463
Random	-	5	0.028	0.031	0.001
	-	7	0.026	0.037	-0.004
	-	9	0.029	0.063	0.004

TABLE 6.1: Summary of Results




			
True Annotations	Tree, Bush, Sidewalk	Temple, Sky	Flower, Bush, Tree, Sidewalk, Building
Empirical Annotations	Tree, Building, People, Bush, Grass	Tree, Building, People, Bush, Grass	Tree, Building, People, Bush, Grass
Vector-Space Annotations	Tree, Bush	Tree, Pole, Grass, Sidewalk, Building, People, Clear Sky	Flower, Bush, Tree, Building, Partially Cloudy Sky
LSI Annotations	Tree, Bush, Grass, Sidewalk	Steps, Wall	Flower, Bush, Tree, Ground

FIGURE 6.3: Example Annotations

Figure 6.2 shows the precision-recall curves for each of the annotation methods and the results are summarised in Table 6.1. The precision-recall curves for the LSI and Vector Space models were generated by increasing the number of images considered for the annotation propagation, M . As would be expected, as M increases, recall also increases due to the increasing number of correctly predicted terms, but precision decreases due to the increased number of incorrect predictions. The curves for the random and frequency distribution based methods were generated by choosing increasing numbers of keywords for annotation. Figure 6.3 shows some example images together with their true and estimated annotations.

The results clearly show that auto-annotation by simple keyword propagation outperforms choosing labels by choosing words based on the frequency distribution of terms. In addition, the LSI based model marginally outperforms the straight vector-space model

in terms of average performance over a number of runs with different training sets. However, from these results, it is not possible to say conclusively that the LSI approach will outperform the vector-space approach in all cases. As in the previous chapters, LSI does have a slight advantage in that it does reduce the dimensionality of the search space dramatically, thus speeding the querying process.

6.2 Using linear-algebra to associate images and terms

Berry et al. (1994) described how Latent Semantic Indexing can be used for cross-language retrieval because it ignores both syntax and explicit semantics in the documents being indexed. In particular, Berry et al. cites the work of Landauer and Littman (1990) who demonstrate a system based on LSI for performing text searching on a set of French and English documents where the queries could be in either French or English (or conceivably both), and the system would return documents in both languages which corresponded to the query. The work of Landauer and Littman negates the need for explicit translations of all the English documents into French; instead, the system was trained on a set of English documents and versions of the documents translated into French, and through a process called ‘folding-in’, the remaining English documents were indexed without the need for explicit translations. This idea has become known as *Cross-Language Latent Semantic Indexing* (CL-LSI).

Monay and Gatica-Perez (2003) attempted to use straight LSI (without ‘folding-in’) with simple cross-domain vectors for auto-annotation. They first created a training matrix of cross-domain vectors and applied LSI. By querying the left-hand subspace they were able to rank an un-annotated query document against each annotation term in order to assess likely annotations to apply to the image. Our approach, described below, is different because we do not explicitly annotate images, but rather just place them in a semantic-space which can be queried by keyword.

Our idea is based on a generalisation of CL-LSI. In general, any document (be it text, image, or even video) can be described by a series of observations made about its content. We refer to each of these observations as terms. The previous chapters introduced the use of ‘visual’ term observations, and the background chapter introduced the idea of observing word occurrences in text documents. There is nothing stopping a term vector having terms from a number of different modalities. For example a term vector could contain term-occurrence information for both ‘visual’ terms and textual annotation terms.

Given a corpus of n documents, it is possible to form a matrix of m observations or measurements (i.e. a term-document matrix). This $m \times n$ observation matrix, \mathbf{O} , essentially represents a combination of terms and documents, and can be factored into

a separate term matrix \mathbf{T} and document matrix \mathbf{D} :

$$\mathbf{O} = \mathbf{T}\mathbf{D}. \quad (6.4)$$

These two matrices can be seen to represent the structure of a semantic-space co-inhabited by both terms and documents. Similar documents and/or terms in this space share similar locations. The advantage of this approach is that it doesn't require *a priori* knowledge and makes no assumptions of either the relationships between terms or documents. The primary tool in this factorisation is the Singular Value Decomposition. This factorisation approach to decomposing a measurement matrix has been used before in computer vision; Tomasi and Kanade (1992) developed an approach, which has become known as Tomasi-Kanade Factorisation, to factoring 3D-shape and motion from measurements of tracked 2D points in image streams.

Our approach consists of two steps. In the first step, a fully-observed *training* observation matrix is created and decomposed into separate term and document matrices. For example, the observations may consist of both 'visual' terms and annotations from a set of training images. The second step consists of assembling an observation matrix for the documents which are to be indexed. These documents need not be fully observed, for example, they may consist of only 'visual' terms. Any unobserved terms are represented by zeros. The document-space of this second observation matrix is then created using the term matrix from the first stage as a basis. The idea behind this is that any term-term relationships that were uncovered in the training stage will be applied to the test data, thus giving the test data *pseudo*-values for the unobserved terms. The net result is that we are left with a new document-space which can be searched by any of the terms used in the training, even if they were not directly observed in the test set.

6.2.1 Decomposing the Observation Matrix

Following the reasoning of Tomasi and Kanade (1992), although modified to fit measurements of terms in documents, we first show how the observation matrix can be decomposed into separate term and document matrices.

Lemma 6.1 (The rank principle for a noise-free term-document matrix). *Without noise, the observation matrix, \mathbf{O} , has a rank at most equal to the number of independent documents or terms observed.*

The rank principle expresses the simple fact that if all of the observed terms are independent, then the rank of the observation matrix would be equal to the number of terms, m . In practice, however, terms are often highly dependent on each other, and the rank is much less than m . Even terms from different modalities may be interdependent;

for example a term representing the colour *red*, and the word “Red”. This fact is what we intend to exploit.

In reality, the observation term-document matrix is not at all noise free. As described in the earlier chapters, the observation matrix, \mathbf{O} can be decomposed using SVD (Golub and Reinsch, 1971) into a $m \times r$ matrix \mathbf{U} , a $r \times r$ diagonal matrix $\mathbf{\Sigma}$ and a $r \times n$ matrix \mathbf{V}^T ,

$$\mathbf{O} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad (6.5)$$

such that $\mathbf{U}^T\mathbf{U} = \mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathcal{I}$, where \mathcal{I} is the identity matrix.

We now partition the \mathbf{U} , $\mathbf{\Sigma}$ and \mathbf{V}^T matrices as follows:

$$\begin{aligned} \mathbf{U} &= \left[\underbrace{\mathbf{U}_k}_k \mid \underbrace{\mathbf{U}_N}_{r-k} \right] \}_{m} \\ \mathbf{\Sigma} &= \left[\begin{array}{c|c} \underbrace{\mathbf{\Sigma}_k}_k & 0 \\ \hline 0 & \underbrace{\mathbf{\Sigma}_N}_{r-k} \end{array} \right] \}_{r-k} \\ \mathbf{V}^T &= \underbrace{\left[\begin{array}{c} \mathbf{V}_k^T \\ \mathbf{V}_N^T \end{array} \right]}_n \}_{r-k} \end{aligned} \quad (6.6)$$

we have

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T + \mathbf{U}_N\mathbf{\Sigma}_N\mathbf{V}_N^T.$$

Assume \mathbf{O}^* is the ideal, *noise-free* observation matrix, with k independent terms. The rank principle implies that the singular values of \mathbf{O}^* are at most k . Since the singular values of $\mathbf{\Sigma}$ are in monotonically decreasing order, $\mathbf{\Sigma}_k$ must contain all of the singular values of \mathbf{O}^* . The consequence of this is that $\mathbf{U}_N\mathbf{\Sigma}_N\mathbf{V}_N^T$ must be entirely due to noise, and $\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ is the best possible approximation to \mathbf{O}^* .

Lemma 6.2 (The rank principle for a noisy term-document matrix). *All of the information about the terms and documents in \mathbf{O} is encoded in its k largest singular values together with the corresponding left and right eigenvectors.*

Thus, the best possible approximation to the ideal observation matrix \mathbf{O}^* is given by

$$\mathbf{O}^* = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T. \quad (6.7)$$

We now define the estimated noise-free term matrix, $\hat{\mathbf{T}}$, and document matrix, $\hat{\mathbf{D}}$, to be

$$\hat{\mathbf{T}} \stackrel{\text{def}}{=} \mathbf{U}_k \quad (6.8)$$

$$\hat{\mathbf{D}} \stackrel{\text{def}}{=} \Sigma_k \mathbf{V}_k^T, \quad (6.9)$$

and from Equation 6.4, we can write

$$\hat{\mathbf{O}} = \hat{\mathbf{T}}\hat{\mathbf{D}}, \quad (6.10)$$

where $\hat{\mathbf{O}}$ represents the estimated noise-free observation matrix.

Note that we could have equally chosen $\hat{\mathbf{T}} \stackrel{\text{def}}{=} \mathbf{U}_k \Sigma_k^{1/2}$ and $\hat{\mathbf{D}} \stackrel{\text{def}}{=} \Sigma_k^{1/2} \mathbf{V}_k^T$, however, the former definition is simpler, and requires less computation in the following steps.

6.2.1.1 Interpreting the decomposition

The two vector bases created in the decomposition form an aligned vector-space of terms and documents. The rows of the term matrix create a basis representing a position in the space of each of the observed terms. The columns of the document matrix represent positions of the observed documents in the space. Similar documents and terms share similar locations in the space.

6.2.2 Using the terms as a basis for new documents

Theorem 6.3 (Projection of partially observed measurements). *The term-matrix of a decomposed fully-observed measurement matrix can be used to project a partially observed measurement matrix into a document matrix that encapsulates estimates of the unobserved terms.*

In order to find a method of projecting a partially-observed observation matrix, \mathbf{P} into the basis created by the term matrix, $\hat{\mathbf{T}}$, we need to perform a little algebraic manipulation of Equation 6.10. The underlying assumption of the projection is that if we were to project the original fully-observed observation matrix (i.e. $\mathbf{P} = \hat{\mathbf{O}}$), then we should get the same document basis.

$$\begin{aligned} \mathbf{P} &= \hat{\mathbf{T}}\hat{\mathbf{D}} \\ \therefore \hat{\mathbf{D}} &= \hat{\mathbf{T}}^{-1}\mathbf{P} \\ &= \hat{\mathbf{T}}^T \hat{\mathbf{T}} \hat{\mathbf{T}}^{-1} \mathbf{P} \\ &= \hat{\mathbf{T}}^T \mathbf{P} \end{aligned} \quad (6.11)$$

Therefore, to project a new partially observed measurement matrix into a basis created from a fully observed training matrix, we need only pre-multiply the new observation matrix by the transpose of the training term matrix. The columns of this new document matrix represent the locations in the semantic space of the documents. In order to query the document set for documents relevant to a term, we just need to rank all of the documents based on their position in the space with respect to the position of the query term in the space (the relevant row of the term matrix). The cosine measure is the most commonly used measure for this task.

Sometimes we want to query with multiple terms. In this case, a vector of terms can be created and projected using Equation 6.11. The projected vector can then be compared against the columns of the document matrix.

Thus far, we have ignored the value of k . The rank principle states that k is such that all of the semantic structure of the observation matrix, minus the noise is encoded in the singular values and eigenvectors. k is also the number of independent, un-correlated terms in the observation matrix. In practice, k will vary across data-sets, and so we have to estimate its value empirically. In section 6.2.5 we show how we choose a value of k , such that the mean-average-precision of a retrieval experiment is maximised.

6.2.3 Summary

In summary, we propose a method of *learning* the semantic structure between terms in a training set, and then applying that structure to a test set. The document space created by this method is unique in that it allows documents to be ranked on their relevance to terms that may not have been explicitly observed within the document, even though the document is relevant to the term.

6.2.4 A Simple Example

Consider two annotated images; I_1 containing the colours red and green and labelled “RED, GREEN”, and I_2 coloured green and blue with the label “GREEN, BLUE”. Suppose that the two images are represented by their dominant colours in RGB-space, and that a visual-vocabulary exists that maps the RGB-space to visual terms. Assume that the $(255, 0, 0)$ RGB triple maps to visual term V_1 , $(0, 255, 0)$ maps to V_2 and $(0, 0, 255)$ to V_3 . Also assume that the annotation terms map to a vocabulary such that “RED” maps to A_1 , “GREEN” to A_2 and “BLUE” to A_3 .

The images and their annotations can be represented by combined *cross-domain* word occurrence vectors by arranging the annotation- and visual-term counts in a vector $(V_1, V_2, V_3, A_1, A_2, A_3)$. The vectors can be arranged in a fully-observed matrix, $\mathbf{O}_{(\text{TRAIN})}$,

$$\mathbf{O}_{(\text{TRAIN})} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Applying the singular-value decomposition yields,

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.289 & 0.500 \\ -0.577 & 0.000 \\ -0.289 & -0.500 \\ -0.289 & 0.500 \\ -0.577 & 0.000 \\ -0.289 & -0.500 \end{bmatrix} \begin{bmatrix} 2.450 & 0 \\ 0 & 1.414 \end{bmatrix} \begin{bmatrix} -0.707 & -0.707 \\ 0.707 & -0.707 \end{bmatrix}.$$

Because the observation matrix in this case did not have any noise, we can see that there are two independently observed terms, and thus the value of k should be 2. The term and document basis matrices are thus (c.f. Equations 6.8 and 6.9),

$$\begin{aligned} \hat{\mathbf{T}} &= \begin{bmatrix} -0.289 & 0.500 \\ -0.577 & 0.000 \\ -0.289 & -0.500 \\ -0.289 & 0.500 \\ -0.577 & 0.000 \\ -0.289 & -0.500 \end{bmatrix} \\ \hat{\mathbf{D}} &= \begin{bmatrix} -1.735 & -1.735 \\ 1.000 & -1.000 \end{bmatrix}. \end{aligned}$$

If we now observe the visual terms of a red image, I_1 , a green image I_2 and a blue image I_3 , we can create a new observation matrix, $\mathbf{P}_{(\text{TEST})}$. The unobserved annotations are set to zeros.

$$\mathbf{P}_{(\text{TEST})} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This new observation matrix can be projected into the semantic space using Equation 6.11,

$$\hat{\mathbf{D}}_{(\text{TEST})} = \begin{bmatrix} -0.289 & -0.577 & -0.289 \\ 0.500 & 0.000 & -0.500 \end{bmatrix}.$$

Now, querying with the textual terms “RED”, “GREEN” and “BLUE” (the 4th, 5th and 6th rows of \mathbf{T} , respectively), gives us the cosine distance to each image:

Image	Cosine similarity with query:		
	“RED”	“GREEN”	“BLUE”
I_1	1.0	0.5	-0.5
I_2	0.5	1.0	0.5
I_3	-0.5	0.5	1.0

This clearly shows that despite the fact the images were un-annotated, they respond correctly to querying by textual terms. The next section illustrates the technique using real images.

6.2.5 Some real examples

In this section, we present experiments using real images from both the Washington and Corel data-sets. Because all of the images in these data-sets have ground truth annotations, it is possible to automatically assess the performance of the retrieval. By splitting the data-sets into a training set and testing set, it is possible to attempt retrieval for each of the annotation terms and mark test images as relevant if they contained the query term in their annotations. Results from using this technique are presented against results using the ‘hard’ annotations from the semantic propagation technique in the previous section.

6.2.5.1 Building a training observation matrix

The process of building the training observation matrix is simple. Basically, as shown in Figure 6.4, vectors for each image are created by appending observations of ‘visual’ and annotation term occurrences. These vectors can then be assembled side-by-side into a matrix.

Although not shown in Figure 6.4, it is possible that the visual term observations could come from any form of descriptor, not just quantised local descriptors. For example, as shown later, it is possible to create observation vectors by combining values from a global colour histogram with annotation term occurrences.

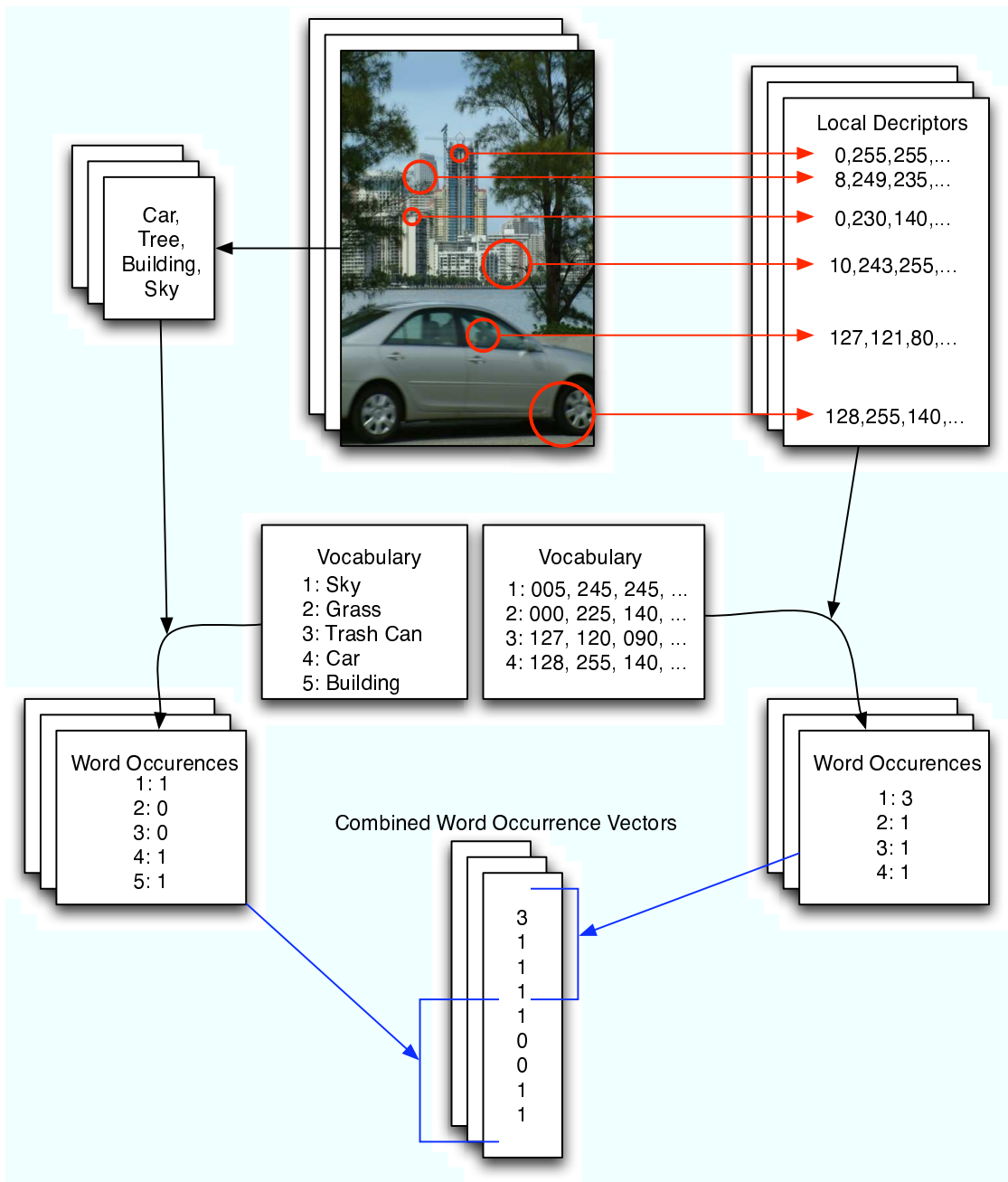


FIGURE 6.4: Generating cross-language vectors of occurrences of 'visual' and annotation terms from a set of annotated images.

6.2.5.2 Experiments with the Washington data-set and SIFT ‘visual’ terms

We split the Washington data-set into a training set of 349 images, and a test set of 348 images. As in earlier chapters of this thesis, each of the images was indexed using ‘visual’ terms from quantised local SIFT descriptors about interest points picked from peaks in a difference-of-Gaussian pyramid. The size of the visual vocabulary was fixed to 3000 terms.

Choosing a good value for k . In order to select a value for k , we need to try and optimise the retrieval. A good statistic of overall retrieval performance is the Mean Average Precision (MAP) (see 2.1.3.2). Plots of the average precision versus varying values of k for four different queries are shown in Figure 6.5. A plot of the MAP over all possible queries, is shown in Figure 6.6.

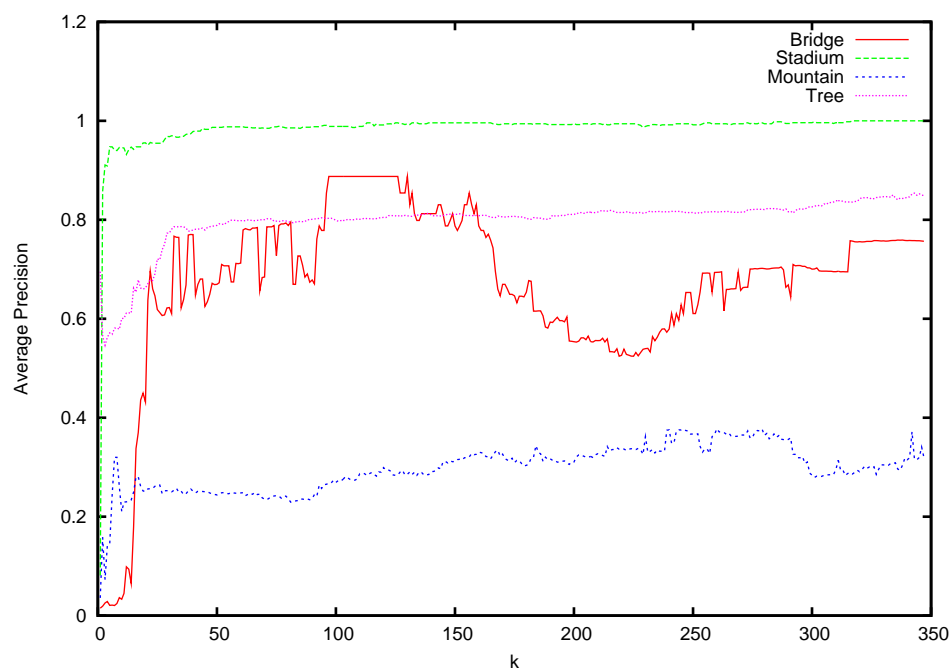
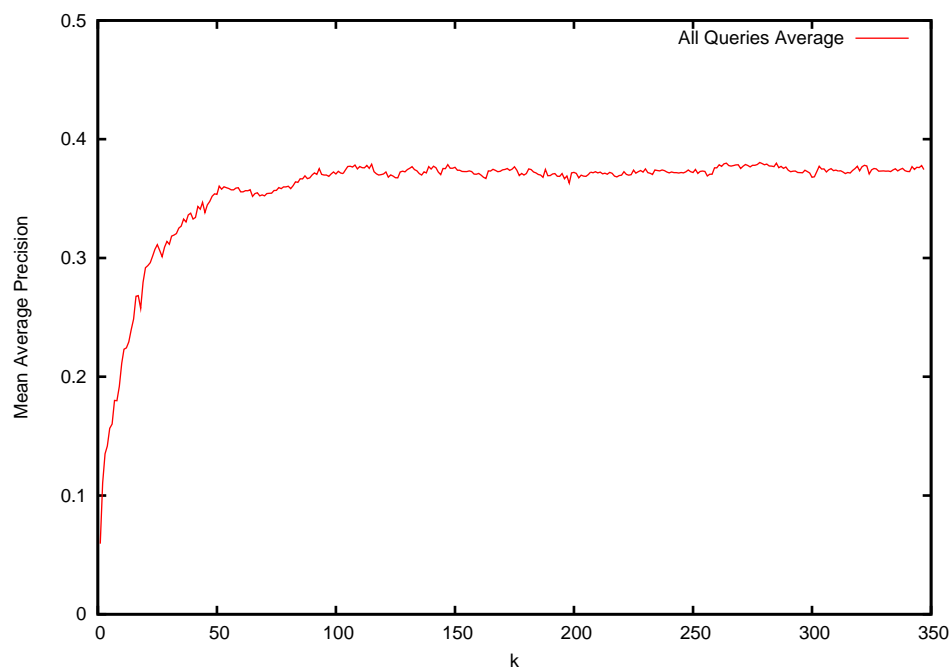
Figure 6.5 shows that there is a very large variation of average precision across different queries. This is in a large part due to biases in both the training set of images and in the test set. For example, both the training set and test set contain an approximately equal number of images of a football stadium, however, the number of *stadium* images in the training set is quite large in comparison to many of the other queries. The net effect is that the “Stadium” query is particularly well trained. Well trained queries can also result from few training images when the training image is sufficiently visually dissimilar to the other images (i.e. it contains a fairly unique combination of visual terms).

Unfortunately, Figure 6.6 doesn’t show a peak from which to select a good value of k , instead it is asymptotic to a mean average precision of about 0.38. However, given the constraint that we want to choose k such that it is the smallest it can be whilst still giving good retrieval, we chose a value of $k = 100$ for the following experiments.

Overall Retrieval Effectiveness. The overall retrieval effectiveness of the technique is characterised in Figure 6.7. As can be seen, the factorisation approach outperforms both the propagation approach at all values of recall. The choice of images for training and test sets is such that the vector-space propagation approach marginally outperforms the LSI propagation approach.

The precision-recall curves in Figure 6.7 don’t truly reflect the whole performance of the approach because certain queries are better performing than others. Figure 6.8(a) illustrates this by showing the average precision for each of the queries, sorted by decreasing precision. Figure 6.8(b) is the same as 6.8(a), but only shows the histogram of average precision for the queries with an average precision of above 0.5.

In order to assess the performance of the factorisation approach to the vector-space propagation approach, Figure 6.9 shows precision histograms for the two algorithms;

FIGURE 6.5: The effect of k on average precision for four different queries.FIGURE 6.6: The effect of k on the Mean-Average Precision over all 170 queries.

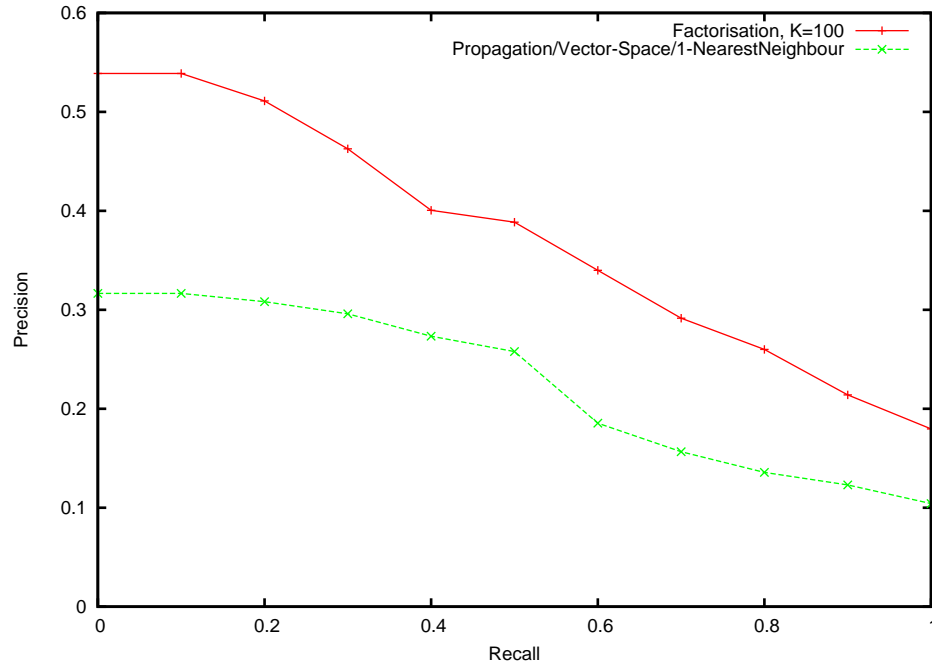


FIGURE 6.7: Average precision-recall curves for the different algorithms over all queries.

6.9(a) shows relative R-Precision of all queries, and 6.9(b) shows the same histogram, but for the queries showing the most difference in performance. On average, the factorisation approach performs better than the propagation approach, although there are a few query terms where the vector-space propagation approach performs slightly better.

Example: Querying for “Bridge”. We now take an example query using the term “Bridge” to investigate the performances of the approaches in more detail. There are ten occurrences of the annotation keyword “Bridge” in the Washington data-set. Of these ten occurrences, four images are in the test set and six in the training set. One of the training images has been labelled with “Bridge”, although it doesn’t actually appear to contain a bridge. This mislabelling of images corresponds to noise, and the algorithms need to be robust to noise within the data-set. The training images are shown in Figure 6.10. Figure 6.11 illustrates the effect on precision over different recall values using both the Factorisation algorithm and the vector-space propagation algorithm. Three different values of k for the factorisation algorithm are shown in the figure. The precision recall curves show that both of the algorithms exhibit perfect precision up to recall values of 0.5, but then tend to drop off.

Figure 6.12 shows the test images containing the “Bridge” keyword, along with the rank-position of the images using the Factorisation and Vector-Space Propagation techniques. The images were retrieved in the same order by the two algorithms, however, the positions at which they occur varies greatly. The factorisation approach retrieved all

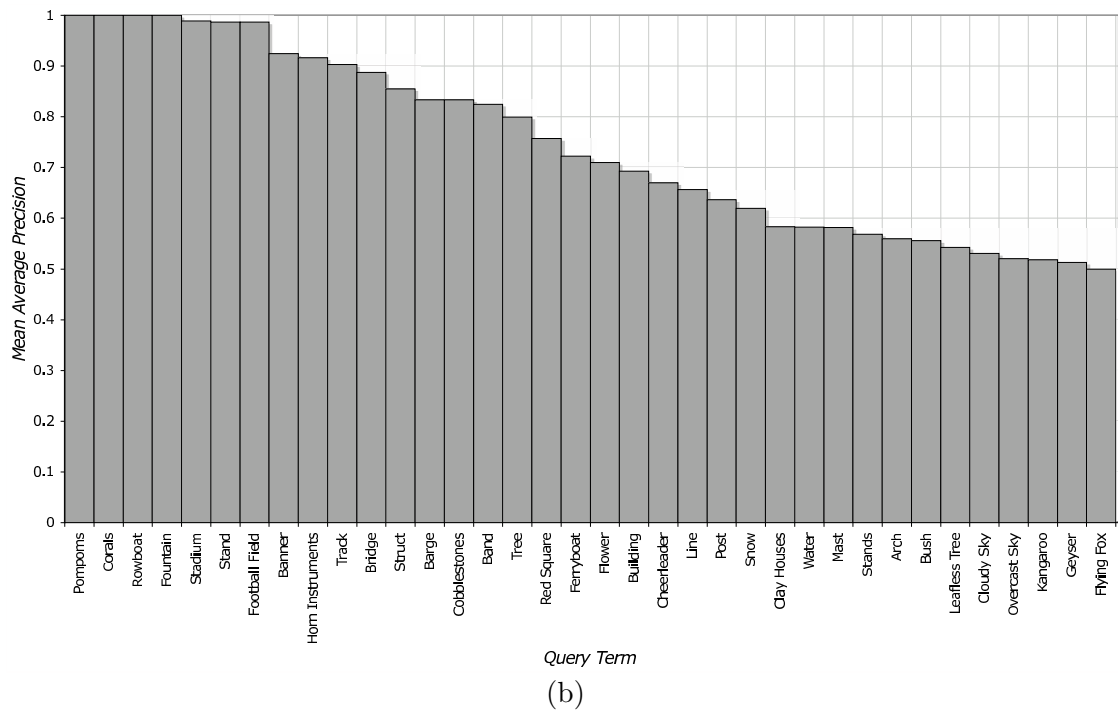
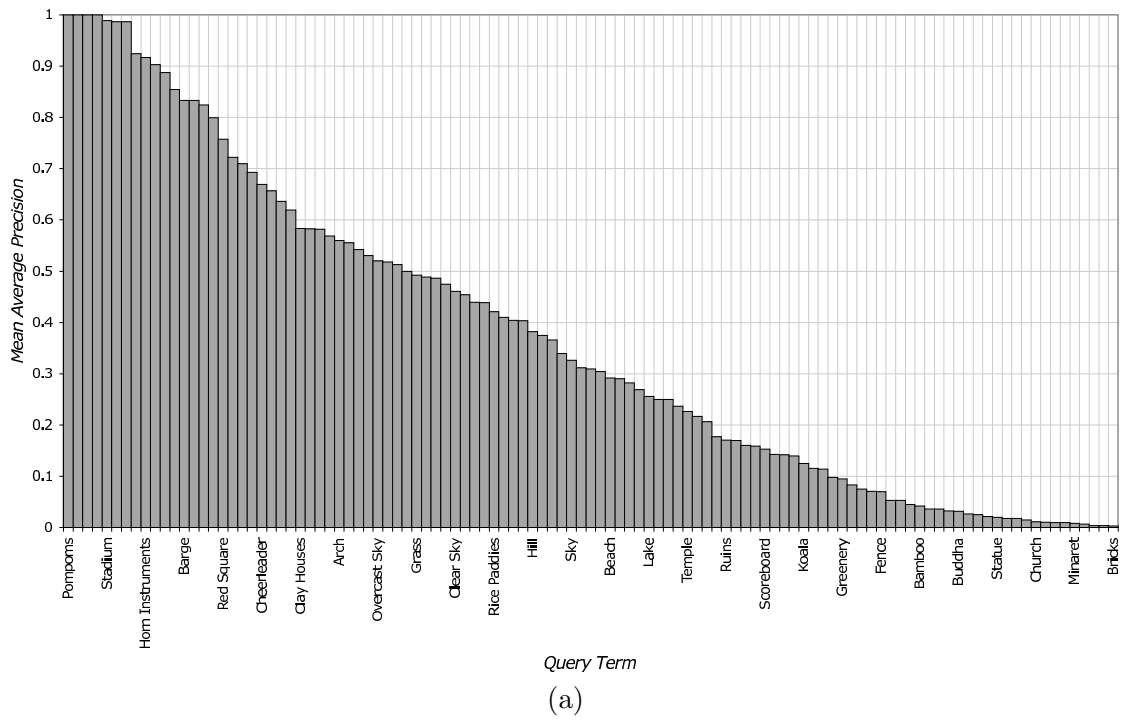


FIGURE 6.8: (a) Average precision of all queries sorted by decreasing precision; (b) Average precision of all queries sorted by decreasing precision of the queries with an average precision of above 0.5.

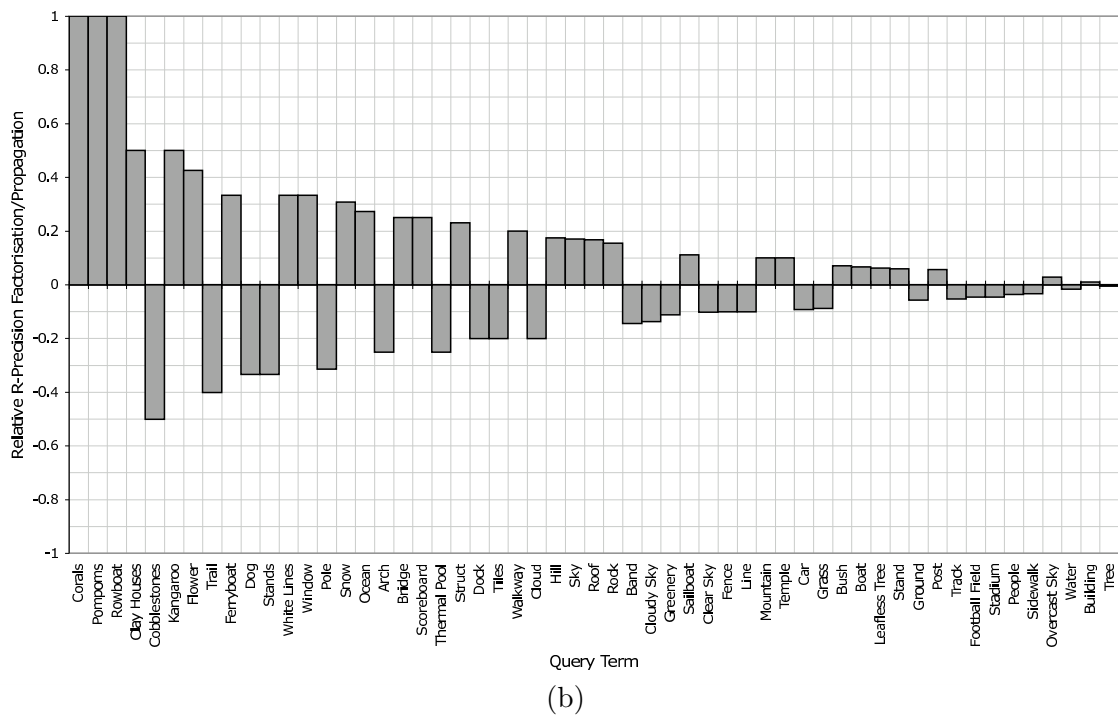
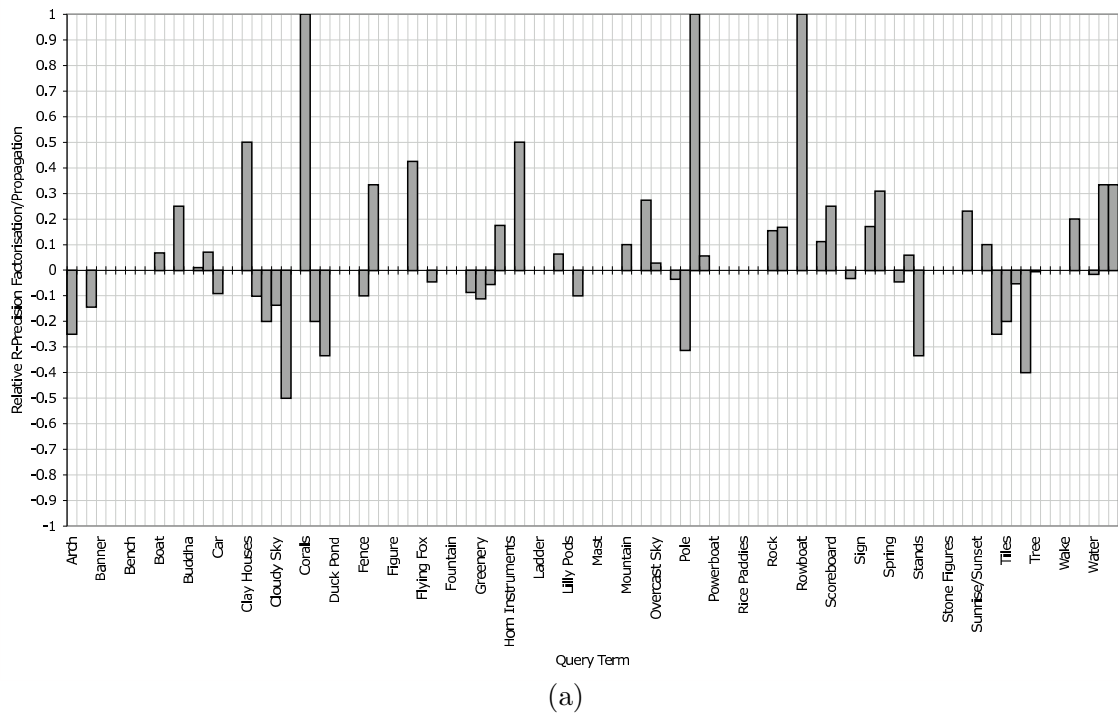


FIGURE 6.9: (a) Relative R-Precision histogram of the Factorisation approach against the Vector-Space approach over all terms. (b) Precision histogram as in (a), but showing only terms with differing performances, ranked by decreasing absolute relative precision. Upward bars indicate better performance for the Factorisation approach, whilst downward bars indicate better performance for the Vector-Space Propagation approach.

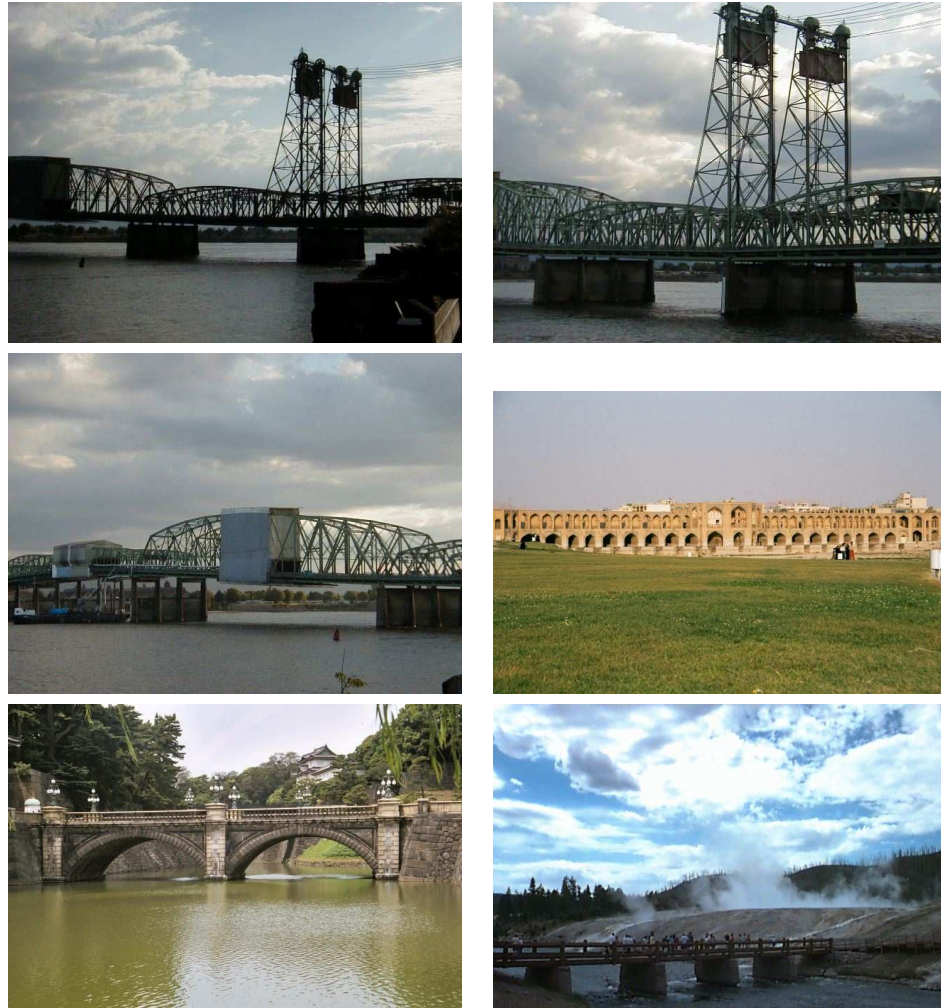


FIGURE 6.10: Training images containing the “Bridge” keyword.

four relevant images within the top five images, whilst the propagation approach didn’t achieve full recall until 332 images had been retrieved.

6.2.5.3 The effect of including colour features in the Washington data-set

We repeated the above experiments using the ‘visual’ terms from the colour descriptor, and with ‘visual’ terms from combining the colour descriptor with the SIFT descriptor. In the case of the colour descriptor alone, an optimal k value set found to be 42, and with the combined terms, k was set to 165.

Using the ‘visual’ terms from the local colour descriptor alone leads to fairly poor retrieval compared to using the SIFT ‘visual’ terms, as shown in Figure 6.13. This is a fairly intuitive result because whilst some of the annotations may have been associated with particular colours, most of them could actually be a range of different colours (i.e. “Tree” is generally green, but “Car” could be green, blue, or any other colour imaginable). Biases in the training set could however lead the semantic space to make incorrect

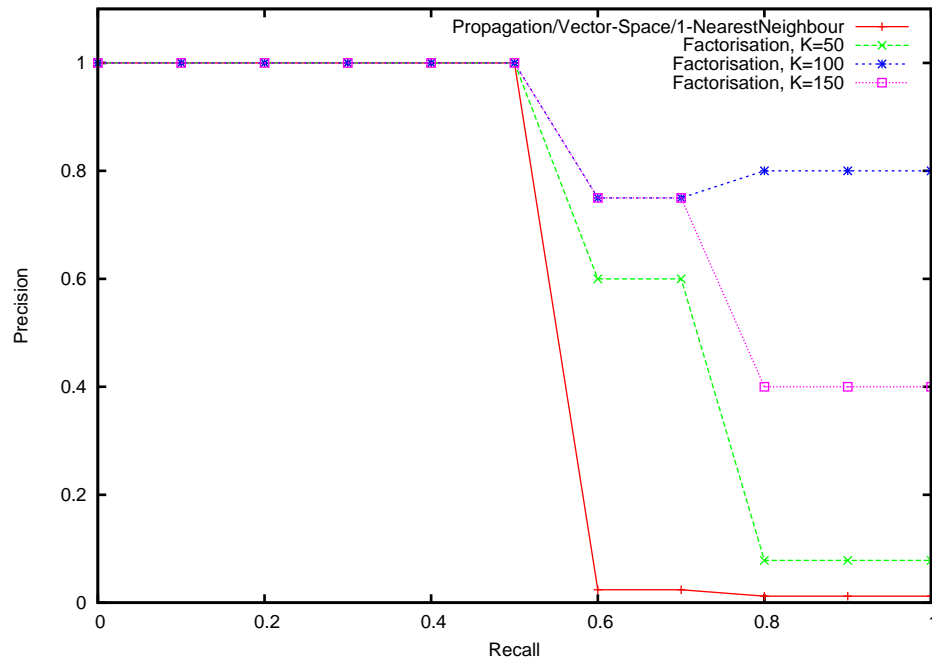


FIGURE 6.11: Precision-Recall curves for querying with the keyword “Bridge” using the Factorisation and Vector-Space Propagation techniques.

associations (e.g. If all the cars were green, the colour green and the terms “Car” and “Tree” would all be close together in the space), which would inherently lead to poor retrieval precision (searches for “Car” may return images of trees as well as images of cars).

Combining the SIFT and colour terms leads to an improvement over colour alone, but does not give an improvement of the average precision compared to the SIFT terms alone. However, that is not to say that the combined colour and SIFT terms don’t help in some queries. Figure 6.14 shows the R-Precision histogram comparing the combined ‘visual’ terms to the SIFT terms. As can be seen, there are a few queries that are marginally improved by including colour information, including some of the annotations that are probably not well characterised by the SIFT descriptor, such as “Clear Sky”, and “Cloudy Sky”.

6.2.5.4 The Corel data-set

Because, as discussed in 4.4.4, the ‘visual’ term representations of the Corel images leads to poor content-based retrieval, it is not useful to attempt to use them for retrieval using the factorisation technique. However, we can demonstrate the power of the technique using simple image features. Whilst in the previous sub-section we proposed using visual terms together with annotation terms in the training observation matrix, this is not the only option. The observation matrix could conceivably contain observations of any type of feature; In this case we demonstrate this by combining the global RGB histogram of





Image	Retrieved Rank Position	
	Factorisation ($k=100$)	Vector-Space Propagation (1-NN)
	1	1
	2	2
	3	125
	5	332

FIGURE 6.12: Test Images and the rank-order in which they were retrieved by the two algorithms.

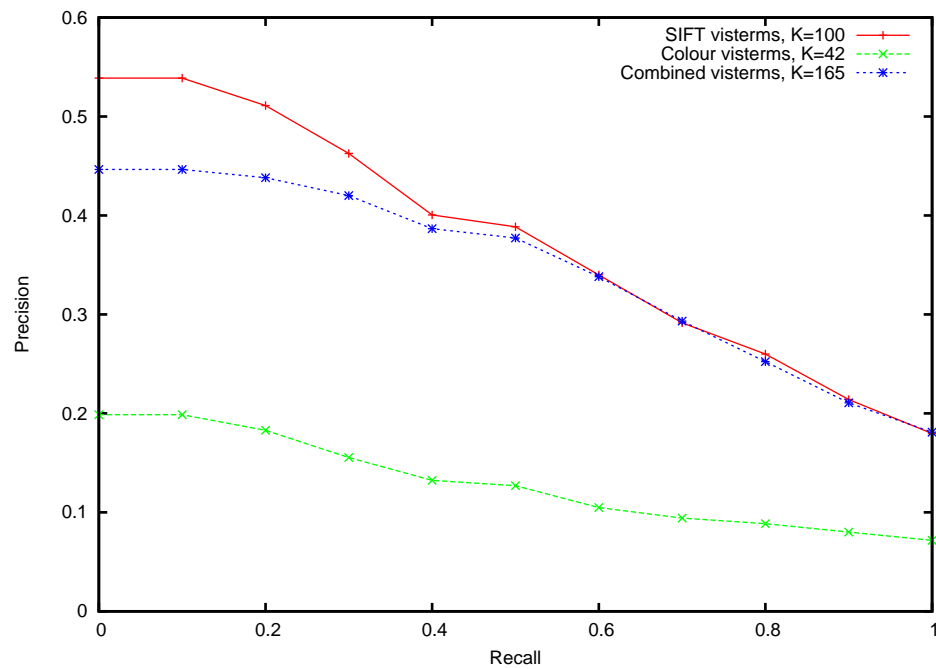


FIGURE 6.13: Precision-Recall curves averaged over all queries with the SIFT ‘visual’ terms, Colour ‘visual’ terms and combined ‘visual’ terms.

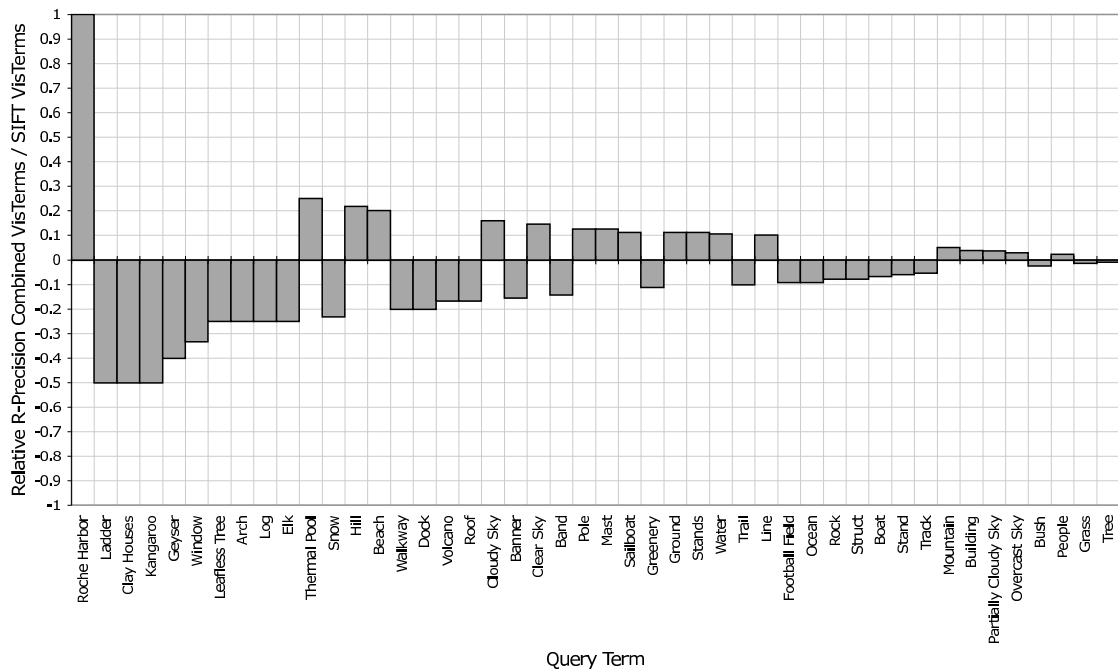


FIGURE 6.14: Relative R-Precision histogram comparing the most differing queries between the use of the combined colour and SIFT ‘visual’ terms against the SIFT ‘visual’ terms alone. Upward bars indicate that the combined terms are better, downward bars show that the SIFT terms are better.

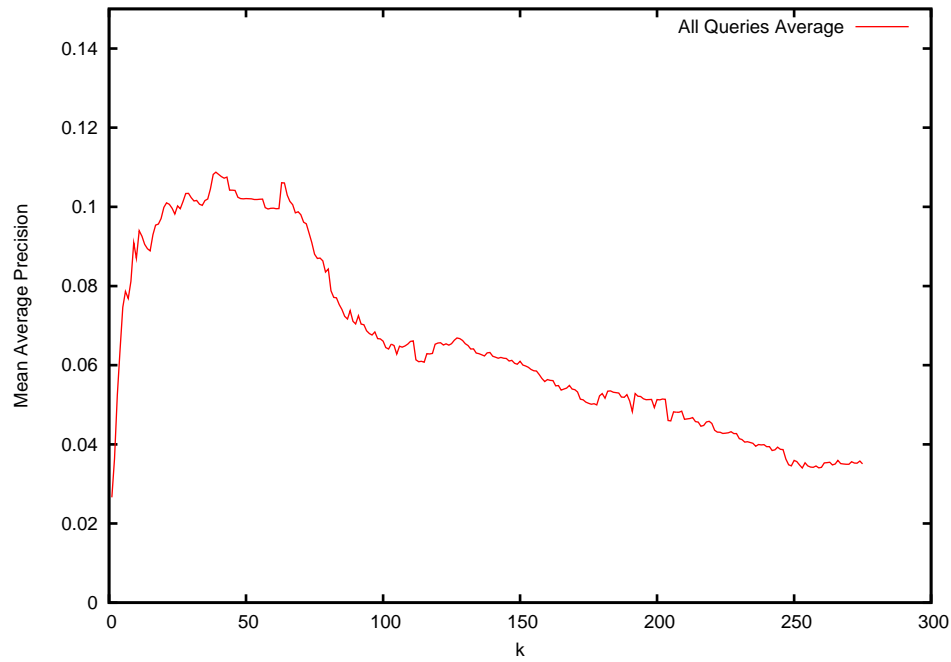


FIGURE 6.15: Plot illustrating the effect of varying k on the mean-average-precision of retrieval with the Corel data-set using RGB-Histogram observations.

each image with its annotation term occurrence vector in order to form the observation matrix. We use the training set of 4500 images and test set of 500 images described by Duygulu et al. (2002).

Figure 6.15 shows the effect of increasing the value of k on the mean-average-precision. From this, k was chosen to be 43. The overall average precision-recall curves of the Factorisation and Vector-Space Propagation approaches are shown in Figure 6.16. As before, the factorisation approach outperforms the propagation approach. Whilst the overall averaged precision-recall curve doesn't achieve a very high recall and falls off fairly rapidly, as before, this isn't indicative of all the queries; some query terms perform much better than others. Figure 6.17 shows histograms of the R-Precision for each query term. Figure 6.18 shows precision-recall curves for some queries with *good* performance.

Ideally, we would like to be able to perform a direct comparison between our factorisation method and the results of the statistical machine-translation model presented by Duygulu et al. (2002), which has become a benchmark against which many auto-annotation systems have been tested. Duygulu et al. present their precision and recall values as single points for each query, based on the number of times the query term was predicted throughout the whole test set. In order to compare results it should be fair to compare the precision of the two methods at the recall given in Duygulu et al.'s results. Table 6.2 summarises the results over the 15 *best* queries found by Duygulu et al. (2002)'s system (base results), corresponding to recall values greater than 0.4.

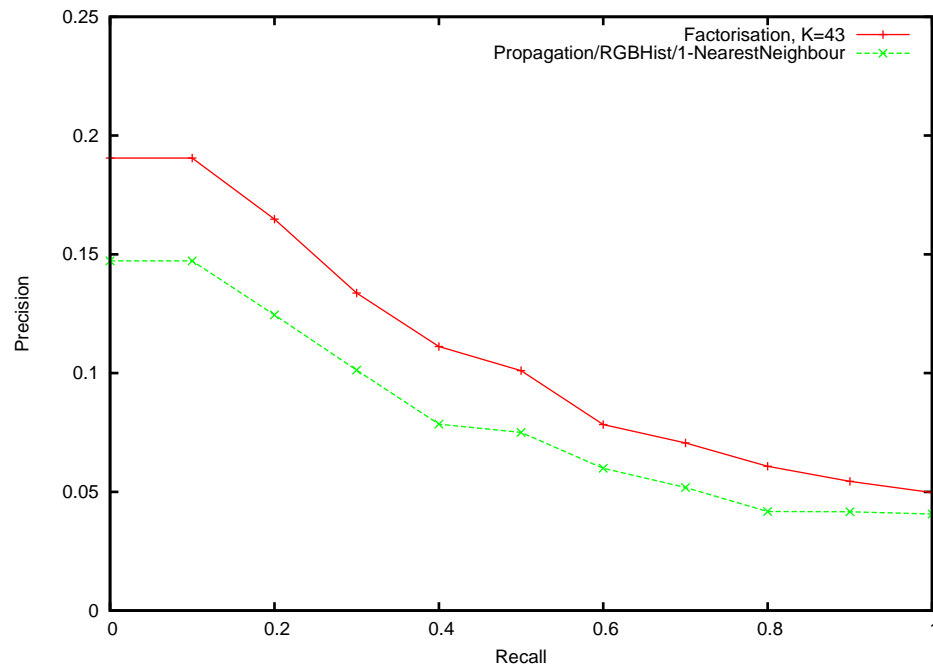


FIGURE 6.16: Average Precision-Recall plots for the Corel data-set using RGB-Histogram descriptors for both the Factorisation and vector-space propagation algorithms.

Query Word	Recall	Precision	
		Machine Translation Base Results, th=0	Factorisation, RGB Histogram, K=43
petals	0.50	1.00	0.13
sky	0.83	0.34	0.35
flowers	0.67	0.21	0.26
horses	0.58	0.27	0.24
foals	0.56	0.29	0.17
mare	0.78	0.23	0.19
tree	0.77	0.20	0.24
people	0.74	0.22	0.29
water	0.74	0.24	0.34
sun	0.70	0.28	0.52
bear	0.59	0.20	0.11
stone	0.48	0.18	0.22
buildings	0.48	0.17	0.25
snow	0.48	0.17	0.54

TABLE 6.2: Comparison of precision values for equal values of recall between Duygulu et al.'s machine translation model and the factorisation approach.

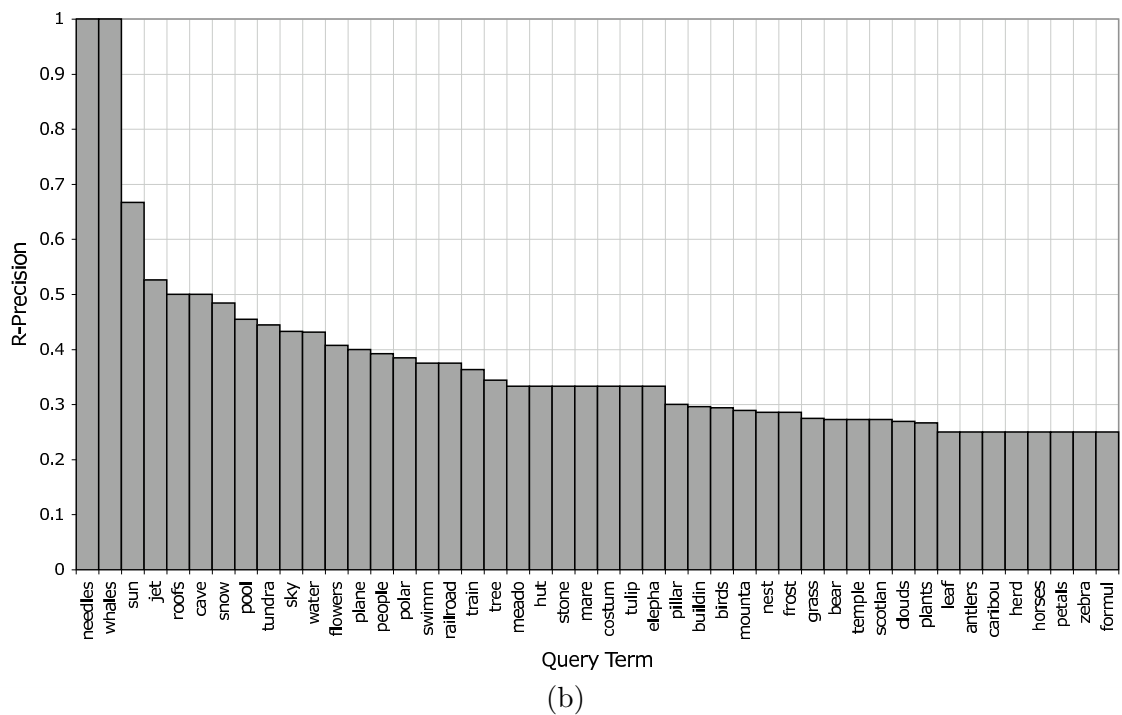
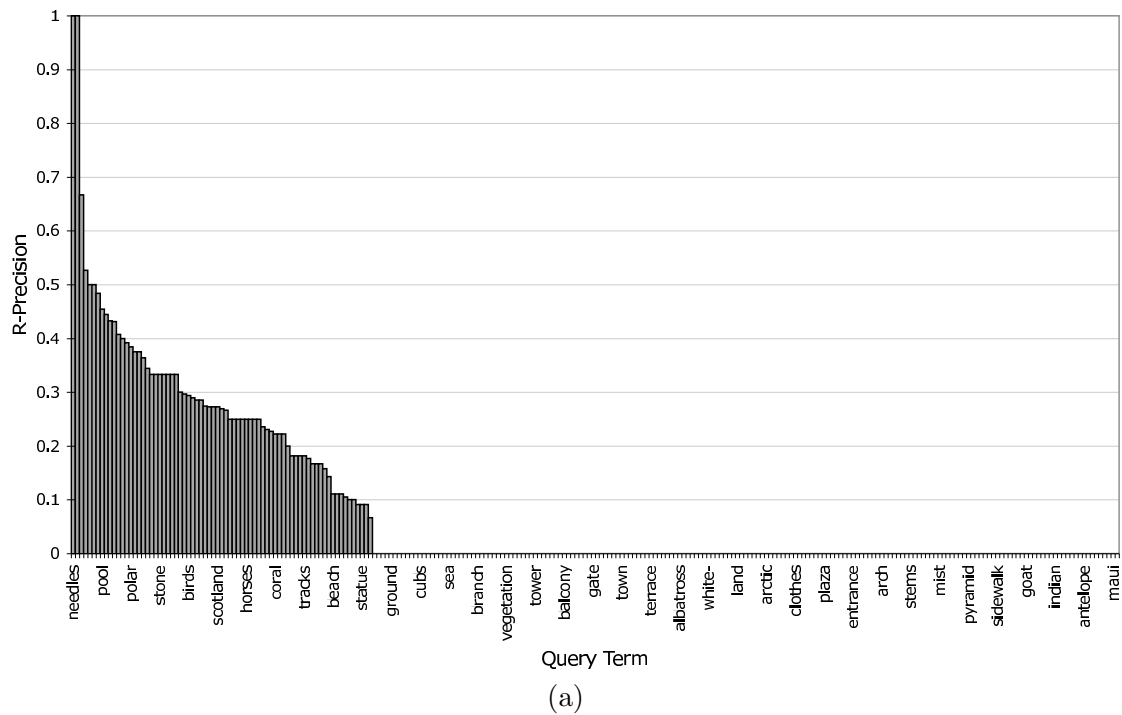


FIGURE 6.17: (a) R-Precision of all queries sorted by decreasing order; (b) R-Precision of all queries with an R-Precision of 0.25 or above, in decreasing order.

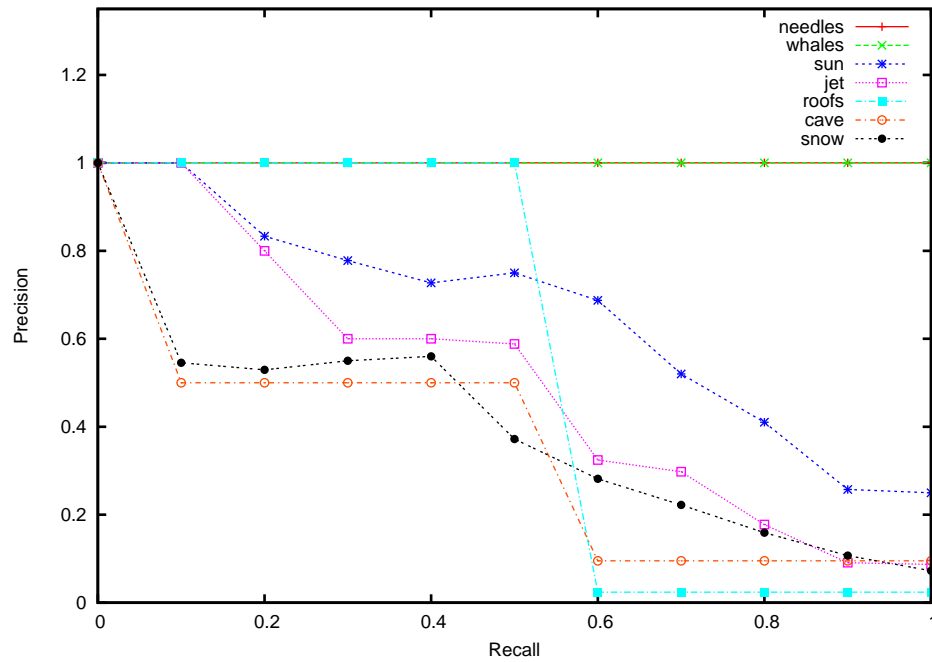


FIGURE 6.18: Precision-Recall curves for the top seven Corel queries using factorisation ($k = 43$).

Table 6.2 shows that nine of the fifteen queries had better precision for the same value of recall with the Factorisation algorithm. This higher precision at the same recall can be interpreted as saying that more relevant images are retrieved with the factorisation algorithm for the same number of images retrieved as with the machine learning approach. This result even holds for Duygulu et al.’s slightly improved *retrained* result set. This implies, somewhat surprisingly, that even by just using the rather simple RGB Histogram to form the visual observations, the factorisation approach performs better than the machine translation approach for a number, of queries. This, however does say something about the relative simplicity of the Corel dataset (Yavlinsky et al., 2005). Because not all of the top performing results (c.f. Figure 6.17) from the factorisation approach are reflected in the *best* results from the machine translation approach, it follows that the factorisation approach may actually perform better on a majority of *good* queries compared to the machine translation model. Of course, whilst the factorisation approach may outperform the machine translation approach in terms of raw retrieval performance, it doesn’t have the capability of applying keywords to individual segmented image regions that the translation model does.

6.2.6 Discussion

The factorisation approach to generating a semantic space for the purpose of performing keyword search on un-annotated image sets described in this section has been shown to perform quite well, even when using a simple global feature such as the RGB histogram.

The approach is exciting because it essentially models the semantic gap in a flexible way, at least as between image descriptors and keywords. The performance of the approach is not the same for all queries - some queries performed really well on the test data, whilst other queries tended to perform less well. The reasons for this are most likely two-fold; firstly, the visual features used to represent the image may not have been sufficient to represent the keyword. Secondly, the training data may not have been sufficient to learn a good representation for the term. In terms of the Corel data-set using RGB histogram features, the factorisation approach works particularly well with annotations that can be described globally across the image by colour alone. For example, searching for ‘sun’ returns images with many warm yellow tones, and searching for ‘snow’ returns images with lots of whitish colours.

The advantage of this technique is that it performs annotation implicitly in a *soft* manner. A *hard* auto-annotator that explicitly applies annotations to images can have problems because it may inadvertently annotate with a similar, but wrong label; for example, labelling an image of a horse with “foal”. Jeon et al. (2003) first noted that this was the case when they compared the retrieval results from a fixed-length hard annotator with a probabilistic annotator. Duygulu et al. (2002) attempt to get around this problem by creating clusters of keywords with similar meaning. However, with our factorisation approach this is not necessary; providing the training data is sufficient, a search for “horse” should also return images of both horses and foals because the terms “horse” and “foal” should share similar locations within the semantic space.

A possible drawback of the factorisation approach is that it is somewhat static. The semantic space must be learnt from a training set, but there is no provision to later learn new terms without repeating the whole process over with a new training set. It is however fairly easy to conceive of ways to solve this problem by updating the semantic space. Possible methods for doing this could be adapted from techniques for updating LSI matrices (Berry et al., 1994).

Much more experimentation needs to be performed to investigate the performance of the factorisation approach. In particular, it would be interesting to use the image descriptors created by Duygulu et al. (2002) (segmented blobs with feature vectors describing their colour, shape, texture, etc) to build our observation matrix, and then to directly compare retrieval results with the CMRM model of Jeon et al. (2003) and the CRM model of Lavrenko et al. (2004). It would also be interesting to see how these models cope with the more general feature observations (such as the global RGB histogram) that the factorisation algorithm permits.

6.3 Summary

This chapter has shown the development of two techniques for enabling keyword search of un-annotated image collections. The first technique works by automatically annotating un-annotated images by propagating the keywords of visually similar images. The second technique performs a linearly-algebraic decomposition of a matrix of observations in an attempt to learn the underlying structure that links visual observations with observations of keyword occurrences. The factorisation approach was shown to outperform the propagation approach over two different data-sets with a range of different visual features.

Chapter 7

Conclusions

“The world is full of obvious things which nobody by any chance ever observes.”

SIR ARTHUR CONAN DOYLE, THE HOUND OF THE BASKERVILLES

“I find that a great part of the information I have was acquired by looking up something and finding something else on the way.”

FRANKLIN P. ADAMS

This thesis has demonstrated to the reader a number of techniques for content-based image retrieval, a subject that is becoming increasingly important with the rapidly increasing amount of digital imagery being accumulated by the people and society of the modern world. This final chapter attempts to draw together and summarise the main conclusions of the preceding chapters and suggest avenues for future research following on from the ideas presented here. Finally the chapter ends with a look at the authors' opinions to where the field of content-based image retrieval is heading in the future.

7.1 Summary and Conclusions

Image retrieval is a wide and varied field encompassing many techniques inspired from other disciplines. This diversity is reflected within each chapter of this thesis. Chapters 3 to 6 describe an array of tools and techniques that can be used for content-based retrieval.

The foundation of content-based image retrieval is the computer vision techniques which make up the low-level feature descriptions that are used to describe and compare image content. Chapter 3 discussed some of the issues related to generating consistent

image descriptions in the presence of noise and other image transformations. The techniques described in the chapter used the concept of saliency in order to generate robust descriptions. Chapter 3 described two evaluations of saliency detectors in which the difference-of-Gaussian detector described by Lowe (2004) was compared to a range of other detectors. All of the detectors tested had strengths and weaknesses in different areas, however, the difference-of-Gaussian detector performed well under most of the distortions it was subjected to. From these results, the difference-of-Gaussian detector was adopted for creating the image descriptions for the experimentation elsewhere in this thesis, however, as discussed in Chapter 4 and by Mikolajczyk et al. (2005), better image descriptions would likely be created by combining the results from multiple detectors.

The final section of Chapter 3 discussed a simple scheme for describing the pixel content of a salient region by its dominant colours. This was achieved by clustering the pixels in RGB space using the mean-shift algorithm. The colour descriptors were used with some success in later chapters, although it was found that whether colour information actually helped retrieval was highly dependent on the data-set.

Techniques for exploring the query-by-example retrieval paradigm are discussed in Chapter 4. The first section of the chapter develops a technique for measuring the content-based retrieval performance of annotated image-sets. The technique attempts to estimate the relevance of retrieved images based on the idea that retrieval algorithms should retrieve semantically similar images, that is images with similar annotations. The section also verifies Sebe et al. (2003)'s result that image description using salient regions can produce better retrieval than with global descriptors.

The second half of Chapter 4 discusses and develops the idea of using text retrieval techniques in combination with salient regions and their descriptors. The technique consisted of quantising the descriptors of each salient region into a 'visual' term and then representing each image by a vector of term occurrences. These term occurrence vectors were then used within a vector-space and Latent Semantic Indexing framework. The results from experiments using these techniques showed generally good performance, although they did highlight a few problems. On the whole, the LSI technique produced better maximum precision (at low recall) than the vector-space model, but performed worse overall. The need to combine different salient region detectors was illustrated in the case of the low-resolution Corel data-set, which in contrast to the Washington data-set was poorly represented when using difference-of-Gaussian salient regions.

Chapter 5 described an application of the retrieval techniques described in the latter parts of Chapter 4. The query-by-example paradigm was extended to work on a mobile device in such a manner that the query image was captured by a camera incorporated into the device. Retrieval performance was demonstrated using images from the National Gallery. In order to ensure a correct match, a re-ranking algorithm was developed that

ensured geometric consistency of matching salient regions within the constraints of a planar-homography.

Finally, Chapter 6 discussed two approaches that attempt to bridge the semantic gap. The first approach proposed simply propagating annotations from similar images. This approach works well if the images are well represented by the low-level features that are used to describe and compare their similarity, such as when using SIFT ‘visual’ terms with the Washington data-set. However, a common problem of all *hard* auto-annotators such as this one was brought to light; images are often mislabelled with keywords that have similar visual appearance to the true keywords, such a mislabelling images of ‘horses’ with ‘foal’. In fact this problem not only with automatic annotators, but also with annotations created manually by humans. This mislabelling can create certain problems in terms of image retrieval. The problem can be assuaged somewhat by methods involving clustering of keywords (Duygulu et al., 2002) or by use of thesauri.

Alternative approaches exist that avoid this mislabelling problem. In the past, probabilistic annotations have been used for ranked retrieval and shown to outperform retrieval using *hard* annotations (Jeon et al., 2003). The second half of Chapter 6 of this thesis suggests another alternative by which an elegant, linearly algebraic manipulation of a matrix of keyword and image-feature observations is shown to produce a *semantic space*. The semantic space this factorisation technique creates represents the underlying structure and links between the keywords and visual features. Un-annotated images can be projected into this semantic space and then searched by keyword. Initial experiments using this approach have shown promise; even when using only simple global features the technique outperforms the machine translation approach described by Duygulu et al. (2002) for a number of search terms.

7.1.1 Novel work in this Thesis

A full list of contributions to the image retrieval community made by this thesis was outlined in the introduction. Not all of those contributions represent novel aspects of the research, and so the contributions with novel value associated with them are reaffirmed here.

- Development of a technique for assessing the content-based retrieval performance of a *query-by-example* style algorithm when using annotated image-sets.
- The extension of the query-by-example paradigm to a mobile device.
- Development of a novel retrieval strategy using quantised local descriptors of salient regions within a vector-space framework.
- Demonstration of a simple technique for auto-annotation by propagating semantics.

- Development of a linear-algebraic technique for building a searchable semantic space with un-annotated images in an attempt to bridge the semantic gap.

7.2 Future Work

Whilst this thesis has covered much ground, there is a lot of scope for improvement in the form of future work. In this section some ideas for future research will be discussed in the context of each of the chapters of this thesis.

7.2.1 Image Description using Saliency

Chapter 3 leaves a number of possibilities open. It has been stated in the past that in order to achieve optimal recognition or retrieval using salient regions, the outputs of multiple salient region detectors should be combined (Mikolajczyk et al., 2005). It would be interesting to investigate what combinations of detectors complement each other in different retrieval scenarios. Obviously, there is still scope for more research into salient region detectors, although this is quite a mature field. The author opines that Scale-Saliency algorithms give perhaps the most pleasing results in regions which appear to be perceptually salient, and that it would be interesting to see if an algorithm could be developed that produces similar regions, whilst still being repeatable. The colour descriptor described at the end of the chapter was shown with little in the way of proof of its performance. Much could be done to assess this, and perhaps improve it.

7.2.2 Image Retrieval using Salient Region Descriptors

The vector-space content-based image retrieval algorithm in Chapter 4 may benefit from using an inverted index structure, such as within the system described by Westmacott (2005). A term-level inverted index, storing the spatial locations, and perhaps the scale of each term and corresponding salient region would allow for some interesting retrieval possibilities as it would allow geometric constraints to be considered at the same time as retrieval, rather than as a separate re-ranking stage.

The single feature morphologies used so far for describing the salient regions are most likely insufficient for truly describing the local characteristics of the region, and thus the image as a whole. Instead of using single feature types, multiple features could be combined in order to produce better image descriptions. An attempt at this was made by using local colour information, however, as previously mentioned this can lead to problems in retrieval of objects where colour is not of importance. A better approach would be to attempt to learn what features are needed to represent a particular object

and use these for indexing. However, this would obviously be difficult as it would require a much higher level of semantic understanding of the image content.

The k -means method used in Chapter 4 has a number of disadvantages when it comes to building a vector quantiser for the construction of the ‘visual words’. For example, if a new set of images were to be added to the database, the existing vocabulary may not be sufficient to describe the new data, and a new quantiser would have to be trained — a very computationally expensive operation. Alternative approaches need to be developed that avoid these problems. Some possible ideas include an adaptive form of split and merge hierarchical clustering and neural network approaches. Another problem with the k -means clustering is that it is very difficult to assess how good the clustering is. Statistically, we can measure the distortion and calculate the Schwarz Criterion, however if we attempt to optimise using these we will need to run the clustering algorithm a number of times with different start points. The current approach to generating clusters was to take a fixed number of random samples from the data-set with which to cluster. A number of cluster results were generated, and the one that produced the best retrieval was used. A better approach to selecting random samples would be to use *Latin Hypercube Sampling*, which should ensure that the samples better represent the underlying data distribution.

Another issue with the quantiser is the time it takes to quantise all of the feature vectors in each image. Currently, a linear search has to be performed in which each feature vector is checked against all of the words in the vocabulary to find the closest. This is a very expensive operation, especially since the SIFT feature vectors are 128 dimensional. This problem is very similar to many multidimensional indexing problems, so it should be possible to employ techniques from this field to reduce the complexity of the problem. Standard tree structures (i.e. b-trees, kd -trees, etc) fail to work efficiently in such high dimensional spaces, however, a special m -ary tree structure known as the triangulation trie (Berman, 1994) may perhaps work well in this situation.

Another interesting avenue to explore would be to investigate how collections of salient regions and their associated descriptors could be used to represent ‘visual’ terms. In this case each quantised salient region descriptor would represent a ‘visual’ letter, and the salient regions that make up a particular semantic object within the image, such as a ‘car’ or ‘tree’, would be represented by a collection of ‘visual’ letters. The order of these letters could be an invariant representation of the spatial location of each letter. The obvious difficulty of this approach is that it essentially requires segmentation, which really requires a higher semantic understanding. Also, it is difficult to see how occlusion could be dealt with; occlusion would likely cause letters to be missed from the ‘visual’ words.

7.2.3 Query by Mobile Device

Chapter 5 leaves a number of possibilities open, especially from a systems engineering point of view. The biggest shortcoming of the approach presented in Chapter 5 is that it relies on the objects being matched/retrieved being planar. This could be easily overcome by indexing multiple views of the objects, and/or improving the geometric re-ranking functionality to use something other than a consistent planar-homographic matrix as a criterion, such as by trying to find a consistent epipolar geometry in the form of the Fundamental matrix. These techniques could perhaps be integrated with some of the ideas for future work on retrieval outlined above.

The system described in Chapter 5 hands all of the computational processing from the device to a server. At the current point in time, this is about the only way in which such a system can work because of constraints in the amount of processing power available on the mobile devices of today. However, as time goes on the amount of processing power in such devices is likely to increase, and it may become more feasible to move more and more of the processing to the device, allowing bandwidth reductions in the amount of data that has to be passed between the device and server. For example, there would be savings in bandwidth, and thus monetary cost, if the mobile client only had to send a vector of ‘visual’-term occurrences to the server instead of sending a whole image. This bandwidth saving would perhaps amount to the difference between sending a few tens of kilobytes to a few hundreds of kilobytes, which might not sound like much in the case of a single device, but would soon add up as the number of devices increased. Of course, the short-term cost of higher power devices would be higher. It would be interesting to investigate this issue in more detail to discover where the optimal distribution of computational power would lie in order to minimise monetary costs.

There are many other aspects, such as with human interface design and usability that also need to be researched. However, perhaps the biggest problem of image retrieval on a mobile device is not from technical difficulties, but rather from an industrial and business point of view, where specific use-cases and applications for such a technology would need to be created in order to assure a suitable business benefit and marketability. The museum scenario created in chapter 5 has relatively little interest to device manufacturers because of its relatively low marketability. On the other hand, such a technology does have a certain amount of *wow factor* or *coolness* associated with it, which should not be neglected, especially if the engineering technicalities are inexpensive to overcome.

7.2.4 Auto-Annotation and Advanced Retrieval

The annotation by propagation approach in the final chapter is fairly limited for retrieval purposes because it applies annotations in a *hard* manner as discussed earlier. However, there are some situations where *hard* annotations are desirable, and the method does

provide an approach to do this. It would be interesting to see if the propagation approach could be improved by changing the way that annotations are propagated. Instead of propagating annotations from the closest M images, a different possibility would be to look at the distribution of similar images, and propagate from a variable number of images based on this distribution. For example, if one image is very close, and the rest have a much greater distance, then perhaps it would be better to propagate from just that one image. If on the other hand there was a number of very similar images, then it may be worthwhile considering propagating all of the annotations from these images.

Of all the parts of this thesis, the factorisation approach to building searchable semantic spaces offers perhaps the most interesting avenue for further investigation. As discussed in more detail in the next section, the semantic gap is what most future content-based image retrieval work is likely to be investigating, and the factorisation technique is particularly well aligned with that direction of investigation. The factorisation algorithm essentially creates a mathematical model of the part of the semantic gap between keyword annotations and image features. Whilst this doesn't allow us to fully bridge the semantic gap, it certainly takes us some of the way there. As previously mentioned the approach needs to be compared with some of the state-of-the-art probabilistic auto-annotators such as the CRM model (Lavrenko et al., 2004), the MBRM model of Feng et al. (2004) and nonparametric density estimation approach (using only global features) proposed by Yavlinsky et al. (2005). It would be interesting to try a different data-set such as the Getty data-set proposed by Yavlinsky et al. (2005). It will also be interesting to see how the approach performs with queries consisting of multiple terms.

An intriguing possibility for the factorisation technique would be for it to be used to associate more abstract semantic structures with image features; for example places, dates and events. As an example, the semantic space could show a relationship between visual features from photographs of people wearing bright colours and coloured beads, the place *New Orleans*, the month of *February*, and the event *Mardi Gras*. Un-annotated photos taken in February, showing similar visual features would automatically respond to searches for *Mardi Gras* or *New Orleans*.

7.3 The Future of CBIR

The field of image retrieval is interesting at the current time. The current trends are twofold; firstly many in the community are becoming increasingly aware of the limitations of current retrieval techniques, especially with regards to the queries formulated by searchers. Secondly, much of the traditional work on image retrieval is being replaced instead with work on video retrieval. Part of the reason for this shift is due to the extra data available in video, such as subtitles, which can massively aid semantic retrieval.

Many researchers are citing the semantic gap, both from the community of professional searchers who are frustrated at the inability of existing systems to accommodate their queries, and from researchers in the content-based retrieval community who believe their particular system may in-part bridge the gap. One current problem is that the term *semantic gap* is meant by many to mean slightly different things. This is an urgent topic that needs to be addressed. The gap needs to be formalised and characterised more clearly and exploration needs to be performed to see what is and is not being done to bridge it. Hare et al. (2006) explores this issue in more detail and attempts to address it.

Semantic retrieval is likely to be the new *buzz-word* for the coming years of image retrieval. How full *semantic retrieval* may be achieved remains to be seen, although this thesis has discussed some approaches by the author and other researchers to get a little closer to this goal. In the authors opinion *hard* automatic annotation is not likely to be a useful avenue to better retrieval because of amongst other factors, the peculiarities of human language, as discussed at the end of Chapter 6. However, the *soft* auto-annotation techniques, such as the factorisation approach to building a semantic space discussed in this thesis, and probabilistic approaches discussed elsewhere, do appear to have a future in our attempts to achieve *semantic retrieval*. It will be interesting to see how these retrieval techniques can be combined with ontologies and other similar techniques for relating concepts and semantics.

It is fair to say, however, that computer vision and image description techniques still have a long way to go before we are able to fully bridge the semantic gap. Today's techniques might be able to tell us that a photograph contains a car on a road, a child and a ball. However, we still have a long way to go before the computer can understand the higher-order semantics of the scene in a meaningful way; in this case that the child is chasing the ball into the road in-front of an oncoming car.

Glossary

Dewey Decimal System A numerical system of classifying and arranging books in a library.

Difference-of-Gaussian An edge detection filter closely linked to the Laplacian-of-Gaussian, formed by subtracting two Gaussian distributions with different variances.

DoG See Difference-of-Gaussian.

Entropy See Shannon Entropy.

Fundamental Matrix A matrix encoding all of the geometrical constraints available given two images of a rigid scene.

Hough Transform A technique for recognising patterns by accumulating votes.

Inverted Index An index into a set of documents of the terms in the documents. The index is accessed by some search method. Each index entry gives the term and a list of documents, possibly with locations within the document, where the term occurs.

JPEG An image compression algorithm developed by the Joint Picture Expert Group.

Latent Semantic Analysis See Latent Semantic Indexing

Latent Semantic Indexing An algebraic model of document retrieval based on a singular value decomposition of the vectorial space of index terms.

Latin Hypercube Sampling A statistical method to generate a distribution of plausible collections of parameter values from a multidimensional distribution.

Log-Entropy A statistical technique used to weight how important a word is to a document. The importance increases proportionally to the number of times a word appears in the document but is offset by how common the word is in all of the documents in the collection or corpus.

LSI See Latent Semantic Indexing.

Mean-shift algorithm A simple iterative procedure that shifts each data point to the average of data points in its neighborhood, seeking out the modes of the distribution of data points.

Planar Homography A linear transform between two planes in space.

Precision In information retrieval, the proportion of the number of relevant documents to the number of all documents retrieved.

QBE See Query By Example

Query A request to a search engine or retrieval system for information.

Query by Example A search method for retrieval systems in which the user formulates a query using existing documents or by creating a proxy document.

Rank The level or position at which a document is retrieved.

RANSAC algorithm An algorithm to estimate parameters in a mathematical model from a data-set when the data set contains many outliers.

Recall In information retrieval, the proportion of retrieved documents of all relevant documents available.

Saliency Referring to parts of an image that stand-out in some manner.

Scale Invariant Feature Transform Robust local feature descriptor that is generated from a three dimensional histogram of gradient orientation at different spatial locations.

Schwartz Criterion In clustering, a number that represents the tradeoff between distortion and the number of clusters.

Semantic Gap The lack of coincidence between the information that one can extract from the visual data within an image and the interpretation that the same data has for a user in a given situation.

Shannon Entropy A measure of randomness in a signal.

SIFT See Scale Invariant Feature Transform.

Singular Value Decomposition A widely used technique to decompose a matrix into several component matrices, exposing many of the useful and interesting properties of the original matrix.

Stemming Refers to procedures for automatically removing certain common suffixes, or word endings, (and sometimes prefixes) in order to increase the frequency count for important words, and also in order to find word occurrences when the word form in the text does not match the word form in the query statement.

SVD See Singular Value Decomposition

Term Frequency-Inverse Document Frequency. A statistical technique used to weight how important a word is to a document. The importance increases proportionally to the number of times a word appears in the document but is offset by how common the word is in all of the documents in the collection or corpus.

Term-Document Matrix Matrix whos elements indicate how many times a given term occurs in a given document.

TF-IDF See Term Frequency-Inverse Document Frequency.

Trie An n -ary tree data structure.

Vector Quantisation A quantization technique in which the basic idea is to code values from a multidimensional vector space into values from a discrete subspace of lower dimension.

Vector-Space Model An algebraic model used for information retrieval. It represents natural language documents in a formal manner by the use of vectors in a multidimensional space.

Wavelet Refers to the representation of a signal in terms of a finite length or fast decaying oscillating waveform.

XML-RPC A simple protocol for making remote procedure requests to Internet-based servers.

Zipfian A distribution of probabilities of occurrence that follows Zipf's law.

Zipf's Law The observation that the frequency of use of the n th-most-frequently-used word in any natural language is approximately inversely proportional to n .

Bibliography

- M Addis, M Boniface, S Goodall, P Grimwood, S Kim, P Lewis, K Martinez, and A Stevenson. SCULPTEUR: Towards a New Paradigm for Multimedia Museum Information Handling. In *International Semantic Web Conference (ISWC 2003)*, pages 582–596, Florida, USA, October 2003.
- L. H Armitage and P. G. B Enser. Analysis of user need in image archives. *Journal of Information Sciences*, 23(4):287–299, 1997.
- J. A Aslam, E Yilmaz, and V Pavlu. A geometric interpretation of r-precision and its correlation with average precision. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 573–574, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-034-5.
- K Barnard, P Duygulu, D Forsyth, N de Freitas, D. M Blei, and M. I Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003. ISSN 1533-7928.
- J Barton and T Kindberg. The challenges and opportunities of integrating the physical world and networked systems. Technical Report HPL-2001-18, HP Labs, 2001.
- N Beckmann, H.-P Kriegel, R Schneider, and B Seeger. The r*-tree: an efficient and robust access method for points and rectangles. In *SIGMOD '90: Proceedings of the 1990 ACM SIGMOD international conference on Management of data*, pages 322–331, New York, NY, USA, 1990. ACM Press. ISBN 0-89791-365-5.
- A. P Berman. A new data structure for fast approximate matching. Technical Report TR-94-03-02, University of Washington, 1994.
- M. W Berry, S. T Dumais, and G. W O'Brien. Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270, University of Tennessee, 1994.
- D. M Blei and M. I Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-646-3.

- D. M Blei, A. Y Ng, and M. I Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. ISSN 1533-7928.
- C Carson, S Belongie, H Greenspan, and J Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8):1026–1038, 2002. ISSN 0162-8828.
- M. L Cascia, S Sethi, and S Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *CBAIVL '98: Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, page 24, Washington, DC, USA, 1998. IEEE Computer Society. ISBN 0-8186-8544-1.
- S Chan, K Martinez, P Lewis, C Lahanier, and J Stevenson. Handling sub-image queries in content-based retrieval of high resolution art images. In *ICHIM (2)*, pages 157–163, 2001.
- S. K Chang and A Hsu. Image information systems: Where do we go from here? *IEEE Transactions on Knowledge and Data Engineering*, 4(5):431–442, 1992. ISSN 1041-4347.
- M Charikar, C Chekuri, T Feder, and R Motwani. Incremental clustering and dynamic information retrieval. *SIAM J. Comput.*, 33(6):1417–1440, 2004. ISSN 0097-5397.
- I. J Cox, M. L Miller, T. P Minka, T Papathomas, and P. N Yianilos. The bayesian image retrieval system, pichunter: Theory, implementation and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.
- H Davis, W Hall, I Heath, G Hill, and R Wilkins. Towards an integrated information environment with open hypermedia systems. In *ECHT '92: Proceedings of the ACM conference on Hypertext*, pages 181–190, New York, NY, USA, 1992a. ACM Press. ISBN 0-89791-547-X.
- H. C Davis, W Hall, I Heath, G. J Hill, and R. J Wilkins. Microcosm: An open hypermedia environment for information integration. Technical Report CSTR 92-15, University of Southampton, 1992b.
- S. C Deerwester, S. T Dumais, T. K Landauer, G. W Furnas, and R. A Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- M Dewey. *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*. Project Guttenberg, 1876.
- M. R Dobie, R. H Tansley, D. W Joyce, M. J Weal, P. H Lewis, and W Hall. A flexible architecture for content and concept based multimedia information exploration. In D. J Harper and J. P Eakins, editors, *The Challenge of Image Retrieval*, Newcastle, 1999, February 1999.

- P Duygulu, K Barnard, J. F. G de Freitas, and D. A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112, London, UK, 2002. Springer-Verlag. ISBN 3-540-43748-7.
- J. P Eakins. Design criteria for a shape retrieval system. *Comput. Ind.*, 21(2):167–184, 1993. ISSN 0166-3615.
- J Eakins and M Graham. Content-based image retrieval. Technical Report JTAP-039, JISC, 2000.
- J. P Eakins, J. M Boardman, and M. E Graham. Similarity retrieval of trademark images. *IEEE MultiMedia*, 5(2):53–63, 1998. ISSN 1070-986X.
- C Eckart and G Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.
- P. G. B Enser. Pictorial information retrieval. *Journal of Documentation*, 51(2):126–170, 1995.
- P. G. B Enser, Y Kompatsiaris, N. E O'Connor, A. F Smeaton, and A. W. M Smeulders, editors. *Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings*, volume 3115 of *Lecture Notes in Computer Science*, 2004. Springer. ISBN 3-540-22539-0.
- P. G. B Enser, C. J Sandom, and P. H Lewis. Automatic annotation of images from the practitioner perspective. In Leow et al. (2005), pages 497–506. ISBN 3-540-27858-3.
- M. F. A Fauzi and P. H Lewis. Query by fax for content-based image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 91–99. Springer-Verlag, 2002. ISBN 3-540-43899-8.
- S. L Feng, R Manmatha, and V Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR (2)*, pages 1002–1009, 2004.
- G. D Finlayson, B Schiele, and J. L Crowley. Using colour for image indexing. In *Challenge of Image Retrieval*, February 1998.
- M Flickner, H Sawhney, W Niblack, J Ashley, Q Huang, B Dom, M Gorkani, J Hafner, D Lee, D Petkovic, D Steele, and P Yanker. Query by Image and Video Content: The QBIC System. *IEEE Computer*, 28(9):23–32, 1995.
- S Gilles. *Robust Description and Matching of Images*. PhD thesis, University of Oxford, 1998.
- G Golub and C Reinsch. *Handbook for automatic computation II, linear algebra*. Springer-Verlag, New York, 1971.

- S Goodall, P. H Lewis, K Martinez, P Sinclair, M Addis, C Lahanier, and J Stevenson. Knowledge-based exploration of multimedia museum collections. In *Proceedings of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*, London, U.K., November 2004.
- Google Inc. Google Image Search: Help. http://www.google.com/help/faq_images.html, 2005.
- W. I Grosky and R Zhao. Negotiating the semantic gap: From feature maps to semantic landscapes. *Lecture Notes in Computer Science*, 2234:33, 2001.
- R. M Haralick, K Shanmugam, and I Dinstein. Texture features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, November 1973.
- J. S Hare and P. H Lewis. Scale saliency: Applications in visual matching, tracking and view-based object recognition. In *Distributed Multimedia Systems 2003 / Visual Information Systems 2003*, pages 436–440, Florida International University, Miami, Florida, September 2003. Knowledge Systems Institute.
- J. S Hare and P. H Lewis. Salient regions for query by image content. In Enser et al. (2004), pages 317–325. ISBN 3-540-22539-0.
- J. S Hare and P. H Lewis. Content-based image retrieval using a mobile device as a novel interface. In R. W Lienhart, N Babaguchi, and E. Y Chang, editors, *Proceedings of Storage and Retrieval Methods and Applications for Multimedia 2005*, pages 64–75, San Jose, California, USA, January 2005a. SPIE.
- J. S Hare and P. H Lewis. On image retrieval using salient regions with vector-spaces and latent semantics. In Leow et al. (2005), pages 540–549. ISBN 3-540-27858-3.
- J. S Hare and P. H Lewis. Saliency-based models of image content and their application to auto-annotation by semantic propagation. In *Proceedings of the Second European Semantic Web Conference (ESWC2005)*, Heraklion, Crete, May 2005c.
- J. S Hare, P. H Lewis, P. G. B Enser, and C. J Sandom. Mind the gap. In E. Y Chang, A Hanjalic, and N Sebe, editors, *Multimedia Content Analysis, Management, and Retrieval 2006*, volume 6073, pages 607309–1–607309–12, San Jose, California, USA, January 2006. SPIE.
- C Harris and M Stephens. A combined corner and edge detector. In M. M Mathews, editor, *Proceedings of the 4th ALVEY vision conference*, pages 147–151, University of Manchester, England, 1988.
- A Hiroike, Y Musha, A Sugimoto, and Y Mori. Visualization of information spaces to retrieve and browse image data. In *VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems*, pages 155–162, London, UK, 1999. Springer-Verlag. ISBN 3-540-66079-8.

- L Hollink, A. T Schreiber, B. J Wielinga, and M Worring. Classification of user image descriptions. *Int. J. Hum.-Comput. Stud.*, 61(5):601–626, 2004. ISSN 1071-5819.
- P Howarth and S Rüger. Evaluation of texture features for content-based image retrieval. In Enser et al. (2004), pages 317–325. ISBN 3-540-22539-0.
- B Hu, S Dasmahapatra, P Lewis, and N Shadbolt. Ontology-based medical image annotation with description logics. In *Proceedings of The 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 77–82. IEEE Computer Society Press, 2003.
- M.-K Hu. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8(2):179–187, 1962.
- T Huang, S Mehrotra, and K Ramchandran. Multimedia analysis and retrieval system (MARS) project. In *Proceedings of the 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval*, 1996.
- IBM Corporation. QBIC – IBM’s Query By Image Content. <http://wwwqbic.almaden.ibm.com>, Accessed 10/9/2005.
- J Jeon, V Lavrenko, and R Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR ’03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 119–126, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-646-3.
- J Jeon and R Manmatha. Using maximum entropy for automatic image annotation. In Enser et al. (2004), pages 24–32. ISBN 3-540-22539-0.
- T Kadir. *Scale, Saliency and Scene Description*. PhD thesis, University of Oxford, Department of Engineering Science, Robotics Research Group, University of Oxford, Oxford, UK, 2001.
- T Kadir and M Brady. Saliency, scale and image description. *Int. J. Comput. Vis.*, 45(2):83–105, 2001.
- T Kadir, A Zisserman, and M Brady. An affine invariant salient region detector. In *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic*, May 2004.
- T Kato, T Kurita, N Otsu, and K Hirata. A sketch retrieval method for full color image database-query by visual example. *Pattern Recognition, 1992 . Vol.1. Conference A: Computer Vision and Applications, Proceedings., 11th IAPR International Conference on*, pages 530–533, 1992.
- J. J Koenderink. The structure of images. *Biological Cybernetics*, 50:363–396, 1984.

- A Laine and J Fan. Texture classification by wavelet packet signatures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11):1186–1191, 1993. ISSN 0162-8828.
- T. K Landauer and M. L Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38, UW Centre for the New OED and Text Research, Waterloo, Ontario, Canada, October 1990.
- V Lavrenko, R Manmatha, and J Jeon. A model for learning the semantics of pictures. In S Thrun, L Saul, and B Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- W. K Leow, M. S Lew, T.-S Chua, W.-Y Ma, L Chaisorn, and E. M Bakker, editors. *Image and Video Retrieval, 4th International Conference, CIVR 2005, Singapore, July 20-22, 2005, Proceedings*, volume 3568 of *Lecture Notes in Computer Science*, 2005. Springer. ISBN 3-540-27858-3.
- M. S Lew, N Sebe, and J. P Eakins, editors. *Image and Video Retrieval, International Conference, CIVR 2002, London, UK, July 18-19, 2002, Proceedings*, volume 2383 of *Lecture Notes in Computer Science*, 2002. Springer. ISBN 3-540-43899-8.
- P. H Lewis, H. C Davis, M. R Dobie, and W Hall. Towards multimedia thesaurus support for media-based navigation. In *First International Workshop on Image Databases and Multimedia Search.*, pages 111–118, 1996a.
- P. H Lewis, H. C Davis, S. R Griffiths, W Hall, and R. J Wilkins. Media-based navigation with generic links. In *HYPertext '96: Proceedings of the the seventh ACM conference on Hypertext*, pages 215–223, New York, NY, USA, 1996b. ACM Press. ISBN 0-89791-778-2.
- T Lindeburg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.
- D Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV*, pages 1150–1157, Corfu, 1999.
- D Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, January 2004.
- W. Y Ma and B. S Manjunath. A comparison of wavelet transform features for texture image annotation. In *ICIP '95: Proceedings of the 1995 International Conference on Image Processing (Vol.2)-Volume 2*, page 2256, Washington, DC, USA, 1995. IEEE Computer Society. ISBN 0-8186-7310-9.
- D Marr. *VISION: A computational Investigation into Human Representation and Processing of Visual Information*. W. H. Freeman and Company, 1982. ISBN 0-7167-1567-8.

- J Matas, D Koubaroulis, and J Kittler. Colour image retrieval and object recognition using the multimodal neighbourhood signature. In D Vernon, editor, *Proceedings of the European Conference on Computer Vision*, LNCS vol. 1842, pages 48–64, Berlin, Germany, June 2000. Springer. ISBN 3-540-67685-6.
- J Matas, O Chum, M Urban, and T Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In P. L Rosin and A. D Marshall, editors, *BMVC*. British Machine Vision Association, 2002. ISBN 1-901725-19-7.
- D Metzler and R Manmatha. An inference network approach to image retrieval. In Enser et al. (2004), pages 42–50. ISBN 3-540-22539-0.
- K Mikolajczyk. *Detection of local features invariant to affine transformations*. PhD thesis, Institut National Polytechnique de Grenoble, France, 2002.
- K Mikolajczyk and C Schmid. A performance evaluation of local descriptors. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 257–263, June 2003.
- K Mikolajczyk, T Tuytelaars, C Schmid, A Zisserman, J Matas, F Schaffalitzky, T Kadir, and L. V Gool. A comparison of affine region detectors. *Accepted in International Journal of Computer Vision*, 2005.
- F Monay and D Gatica-Perez. On image auto-annotation with latent space models. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278. ACM Press, 2003. ISBN 1-58113-722-2.
- Y Mori, H Takahashi, and R Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, 1999.
- H Müller, S Marchand-Maillet, and T Pun. The truth about corel - evaluation in image retrieval. In Lew et al. (2002), pages 38–49. ISBN 3-540-43899-8.
- W Niblack, R Barber, W Equitz, M. D Flickner, E. H Glasman, D Petkovic, P Yanker, C Faloutsos, and G Taubin. QBIC project: querying images by content, using color, texture, and shape. In *Storage and Retrieval for Image and Video Databases*, 1993.
- S Obdržálek and J Matas. Local affine frames for image retrieval. In Lew et al. (2002), pages 318–327. ISBN 3-540-43899-8.
- S Obdrzalek and J Matas. Image retrieval using local compact DCT-based representation. In *DAGM-Symposium 2003*, pages 490–497, 2003.
- A Oliva and A Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001. ISSN 0920-5691.

- A Oliva and A. B Torralba. Scene-centered description from spatial envelope properties. In *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 263–272, London, UK, 2002. Springer-Verlag. ISBN 3-540-00174-3.
- S Ornager. Image retrieval: Theoretical and empirical user studies on accessing information in images. In *Proceedings of the 60th American Society of Information Retrieval Annual Meeting*, volume 34, pages 202–211, 1997.
- L Page and S Brin. The anatomy of a search engine. In *The 7th International WWW Conference (WWW 98)*, Brisbane, Australia, April 1998.
- G Pass, R Zabih, and J Miller. Comparing images using color coherence vectors. In *Proceedings of ACM Multimedia*, pages 65–73, 1996.
- M. F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- Y Rui, T Huang, and S Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4): 39–62, April 1999.
- Y Rui, K Chakrabarti, S Mehrotra, Y Zhao, and T. S Huang. Dynamic clustering for optimal retrieval in high dimensional multimedia databases. Technical Report TR-MARS-10-97, Beckmann Insitute, University of Illinois at Urbana-Champaign, 1997a.
- Y Rui, T. S Huang, and S Mehrotra. Content-based image retrieval with relevance feedback in MARS. In *ICIP (2)*, pages 815–818, 1997b.
- Y Rui, T. S Huang, and S Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36, 1998.
- A Salah, E Alpaydin, and L Akarun. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):420–425, 2002.
- G Salton, A Wong, and C. S Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975. ISSN 0001-0782.
- C Schmid and R Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.
- C Schmid, R Mohr, and C Bauckhage. Evaluation of interest detectors. *Int. J. Comput. Vis.*, 37(2):151–172, 2000.

- J. T Schwartz and M Sharir. Identification of partially obscured objects in two and three dimensions by matching noisy characteristic. *Int. J. Rob. Res.*, 6(2):29–44, 1987. ISSN 0278-3649.
- N Sebe, Q Tian, E Louprias, M Lew, and T Huang. Evaluation of salient point techniques. *Image and Vision Computing*, 21:1087–1095, 2003.
- N Sebe and M. S Lew. Comparing salient point detectors. *Pattern Recognition Letters*, 24(1-3):89–96, 2003.
- A Shokoufandeh, I Marsic, and S Dickinson. View-based object recognition using saliency maps. *Image Vis. Comput.*, 17(5-6):445–460, 1999.
- J Sivic and A Zisserman. Video google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, pages 1470–1477, October 2003.
- A. W. M Smeulders, M Worring, S Santini, A Gupta, and R Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000. ISSN 0162-8828.
- J Smith. Image retrieval evaluation. In *IEEE Workshop on Content-based Access of Image and Video Libraries, Santa Barbara, California*, 1998.
- J. R Smith and S.-F Chang. Single color extraction and image query. In *ICIP '95: Proceedings of the 1995 International Conference on Image Processing (Vol. 3)-Volume 3*, page 3528, Washington, DC, USA, 1995. IEEE Computer Society. ISBN 0-8186-7310-9.
- J. R Smith and S.-F Chang. Transform features for texture classification and discrimination in large image databases. In *Proceeding of IEEE International Conference on Image Processing*, volume 3, pages 407–411, Austin, Texas, 1994.
- J. R Smith and S.-F Chang. Tools and techniques for color image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 426–437, 1996.
- M. A Stricker and M Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 381–392, 1995.
- M. J Swain and D. H Ballard. Color indexing. *Int. J. Comput. Vision*, 7(1):11–32, 1991. ISSN 0920-5691.
- H Tamura, S Mori, and T Yamawaki. Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(460-473):6, 1978.
- H Tamura and N Yokoya. Image database systems: A survey. *Pattern Recognition*, 17(1):29–43, 1984.

- C Tomasi and T Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, November 1992.
- T Tuytelaars and L. V Gool. Content-based image retrieval based on local affinity invariant regions. In *Third International Conference on Visual Information Systems*, pages 493–500, 1999.
- University of Washington. Ground truth image database. <http://www.cs.washington.edu/research/imagetdatabase/groundtruth/>, Accessed 6/11/2003.
- C. C Venters and M. D Cooper. A review of content-based image retrieval systems. Technical Report JTAP-054, JISC, 2000.
- E Vincent and R Laganière. Detecting planar homographies in an image pair. In *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis*, pages 182–187, Pula, Croatia, June 2001.
- M Westmacott. *Content Based Image Retrieval: Analogies with text*. PhD thesis, University of Southampton, January 2005.
- M Westmacott and P. H Lewis. An inverted index for image retrieval using colour pair feature terms. In *Proceedings of the SPIE Image and Video Communications and Processing Conference*, pages 881–889, January 2003.
- D. A White and R Jain. Similarity indexing: Algorithms and performance. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 62–73, 1996.
- L Wiscott, J.-M Fellous, N Krüger, and C von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pat. Anal. and Mach. Int.*, 7(19):775–779, 1997.
- H. J Wolfson and I Rigoutsos. Geometric hashing: An overview. *IEEE Comput. Sci. Eng.*, 4(4):10–21, 1997. ISSN 1070-9924.
- A Yavlinsky, E Schofield, and S Rüger. Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation. In W. K Leow, M. S Lew, T.-S Chua, W.-Y Ma, L Chaisorn, and E. M Bakker, editors, *Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR)*, volume 3568 of *Lecture Notes in Computer Science*, pages 507–517, Singapore, July 2005. Springer-Verlag. ISBN 3-540-27858-3.
- C. T Zahn and R. Z Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on Computing*, 21(3), March 1972.
- H Zhang and D Zhong. A scheme for visual feature based image retrieval. In *Proceedings SPIE Storage and Retrieval for Image and Video Databases*, 1995.
- R Zhao and W. I Grosky. From features to semantics: Some preliminary results. In *IEEE International Conference on Multimedia and Expo (II)*, pages 679–682, 2000.